

Lecture Notes in Networks and Systems 1351

Zheng Xu

Saed Alrabaee

Octavio Loyola-González


Nurul Hidayah Ab Rahman *Editors*

Cyber Security Intelligence and Analytics

The 6th International Conference
on Cyber Security Intelligence and
Analytics (CSIA 2024), Volume 1

 Springer

Series Editor

Janusz Kacprzyk , *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*

Okay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*

Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*

Imre J. Rudas, *Óbuda University, Budapest, Hungary*

Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Zheng Xu · Saed Alrabaee ·
Octavio Loyola-González ·
Nurul Hidayah Ab Rahman
Editors


Cyber Security Intelligence and Analytics

The 6th International Conference
on Cyber Security Intelligence and Analytics
(CSIA 2024), Volume 1

Editors

Zheng Xu
School of Computer and Information
Engineering
Shanghai Polytechnic University
Shanghai, China

Octavio Loyola-González
NTT DATA
Madrid, Madrid, Spain

Saed Alrabaee 
United Arab Emirates University
Abu Dhabi, Abu Dhabi, United Arab
Emirates

Nurul Hidayah Ab Rahman
Universiti Tun Hussein Onn Malaysia
Selangor, Johor, Malaysia

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-88286-9

ISBN 978-3-031-88287-6 (eBook)

<https://doi.org/10.1007/978-3-031-88287-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The 6th International Conference on Cyber Security Intelligence and Analytics (CSIA 2024) is an international conference dedicated to promoting novel theoretical and applied research advances in the interdisciplinary agenda of cyber security, particularly focusing on threat intelligence and analytics and countering cybercrime. Cyber security experts, including those in data analytics, incident response and digital forensics, need to be able to rapidly detect, analyze and defend against a diverse range of cyber threats in near real-time conditions. For example, when a significant amount of data is collected from or generated by different security monitoring solutions, intelligent and next generation big data analytical techniques are necessary to mine, interpret and extract knowledge of these (big) data. Cyber threat intelligence and analytics are among the fastest growing interdisciplinary fields of research bringing together researchers from different fields such as digital forensics, political and security studies, criminology, cyber security, big data analytics, machine learning, etc. to detect, contain and mitigate advanced persistent threats and fight against organized cybercrimes. The 6th International Conference on Cyber Security Intelligence and Analytics (CSIA 2024), building on the previous successes meeting in Shanghai, China (2023), (online meeting from 2020 to 2022 due to COVID-19), in Wuhu, China (2019), is proud to be in the 6th consecutive conference year.

We are organizing the CSIA 2024 at the Mulian Hotel Guangzhou Zhujiang New Town, Guangzhou, China. It will feature a technical program of refereed papers selected by the international program committee, a keynote address. Each paper was reviewed by at least two independent experts. The conference would not have been a reality without the contributions of the authors. We sincerely thank all the authors for their valuable contributions. We would like to express our appreciation to all members of the program committee for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

Our special thanks are due also to the editors of the Springer book series “Lecture Notes in Networks and Systems”, Thomas Ditzinger and Praveena Anandhan for their assistance throughout the publication process.

Organization

Steering Committee Chair

Kim-Kwang Raymond Choo	University of Texas at San Antonio, USA
------------------------	---

General Chair

Zheng Xu	Shanghai Polytechnic University, China
----------	--

Program Committee Chairs

Saed Alrabaee	United Arab Emirate University, UAE
Octavio Loyola-González	NTT DATA, Spain
Nurul Hidayah Ab Rahman	Universiti Tun Hussein Onn Malaysia, Malaysia

Publication Chairs

Juan Du	Shanghai University, China
Shunxiang Zhang	Anhui University of Science & Technology, China

Publicity Chairs

Neil. Y. Yen	University of Aizu, Japan
Junchi Yan	Shanghai Jiaotong University, China

Local Organizing Chairs

Ge You	Guangdong Innovative Technical College, China
Jiaqi Wang	Guangdong Innovative Technical College, China
Sulin Pang	Jinan University, China

Contents

Construction of a Comprehensive Safety Guarantee System for College Students Based on Digital Twin Technology	1
<i>Junyan Song</i>	
Data Resource Value Evaluation Algorithm Based on Fuzzy Theory and Catastrophe Series Method	13
<i>Lei Wang</i>	
Comparative Study of ARIMA Model and Long Short Term Memory Network (LSTM) in Economic Management	24
<i>Xiwen Wang</i>	
The Application and Empirical Study of Causality in the Theory Graph	33
<i>Guijiao He</i>	
Research on Machining and Simulation Optimization System of Automobile Steering Knuckle Based on Advanced Algorithm	43
<i>Xiaopeng Chang, Siyu Chen, Xiyu Zhang, Bangcheng Zhang, and Bo Yu</i>	
Network Partitioning and Demand Characterization for Management of Urban Low Voltage Power Distribution Systems	56
<i>Haisheng Hong, Yongshu Chen, Zhifang Zhu, Zheng Sun, Jiarui Guo, and Qin Lin</i>	
Automation and Data Aggregation Algorithm for Low Code Development in Power Grid Mobile Report Generation	70
<i>Wenting Wei and Jie Zhang</i>	
Big Data Analysis and Smart Grid Security Event Monitoring and Response in the Power Internet of Things	81
<i>Hongyu Ke, Zhaoyu Zhu, Shuo Yang, Yi Tang, Ning Xu, and Xin He</i>	
Construction of an Online Autonomous Learning Model Based on Artificial Intelligence ChatGPT	92
<i>Aimin Li and Chenming Yang</i>	
Application of PMS Graphic Intelligent Recognition and Analysis Based on Contact Diagram Automatic Generation Technology	105
<i>Wei Ma, Qiang Li, Yuan Yao, and Xinkai Chen</i>	

Application of Genetic Algorithm in Reasonableness Evaluation
of Environmental Design Space Layout 115
Xiuliang Xi and Jianmei Wei

Construction of an Intelligent Recommendation Model for Digital Media
Content Based on Artificial Intelligence 127
Xiaoning Tang

Research on Link Selection and Allocation for IoT Localization Systems
Based on an Improved Ant Colony Algorithm 140
Jiong Zhang, Meng Xu, and Liying Wang

Cross-Domain Sharing and Privacy Protection Method of Fused
Multi-source Data in Visual Internet of Things 151
Longjie Zhu, Xuming Fang, Xinlei Yang, and Ming Li

Construction of Cold Chain Logistics and Distribution Site Selection
System Based on Multi-objective Optimization Model 162
Hanjie Jia, Hua Jiang, and Manjiang Chen

Hole Detection Algorithm Based on Channel Fusion Siamese Network 174
*Nuan Sun, Chunhe Shi, Yanchao Cui, Yaran Wang, Xiaoying Shen,
and Xinru Shao*

Intelligent Database Triggers Enable Advanced Analysis of Data Recorded
in Audit Logs 184
Yongna Li, Cuiping Li, and Zhaoxia Cui

Improvement of Principal Component Analysis Algorithm and Its
Simulation Experiment 194
Ling Zhang

Application of Computer Information Technology in Intelligent Analysis
and Decision-Making Support of Diagnosis and Treatment Data 208
Yan Gao and Yinsong Zhang

Security Vulnerability Detection and Defense of Smart Home Systems
Based on the Internet of Things 220
Zhenghui Zhao and Miao Chen

In-Depth Discussion and Thorough Research on High-Availability Data
Technology Within the Cloud Environment 229
Lei Yao

State Estimation and Fault Location of Multi-machine Power System Using Graph Neural Network and Variational Autoencoder	239
<i>Fan Zhang, Mengyan Guo, and Ya Wang</i>	
Fault Detection and Diagnosis of Ship Circuit Based on Machine Learning Algorithm	248
<i>Shuyan Liu</i>	
The Process of Building Color Extraction is Optimized with K-means Clustering Algorithm	256
<i>Jian Liu and Junru Chen</i>	
Research on Risk Monitoring and Early Warning Technology for Special Disaster Emergency Rescue Site Based on Reformer Model	265
<i>Lei Zhang, Yufeng Fan, and Zhenpeng An</i>	
Research on Intelligent Optimal Scheduling Algorithm for Vehicle Exhaust Emission in Railway Transportation System	275
<i>Zixiang Xu, Xiaokai Zhou, and Yishan Wang</i>	
OLAP Technology Financial Statistics Information Platform Based on Big Data Analysis	283
<i>Hanyue Xu</i>	
Productivity Estimation Based on Optical Remote Sensing Image Spatiotemporal Fusion Algorithm	294
<i>Jingyi Chu</i>	
Partial Differential Equation Data Fusion Algorithm Based on D_S Evidence Theory and Fuzzy Mathematics	304
<i>Ximei Shi</i>	
Exploration of the Application of Blockchain Technology in Secure Storage and Sharing of Archival Information	314
<i>Xiaoning Chen</i>	
Efficient Data Classification and Prediction Using Random Forest (RF) and Gradient Boosting Machine (GBM)	325
<i>Junliang Du, Xiaoyi Wang, Junpeng Chen, Ziyang Zhao, and Yang Zheng</i>	
Low Carbon Transformation Effect of Logistics Enterprises Based on Adaboost Regression Algorithm	334
<i>Li Yao</i>	

Data Privacy Protection Technology in Digitalization of Power System
Security Management: Application of Homomorphic Encryption Scheme 346
*Dong Wang, Caihua Liu, Lifei Chen, Junliang Wang, Feng Su,
and Xiangyang Li*

Mobile Communication Network Base Station Deployment Under 5G
Technology: A Discussion on the Combination of Genetic Algorithm
and Machine Learning 357
Moxin Zhang, Yimin Wang, and Bingjiao Shi

Application of Digital Intelligent Algorithm in the Construction of Internet
Cultural Communication Platform 370
Xi Chen

Application of Multi-model Fusion Deep NLP System in Classification
of Brain Tumor Follow-Up Image Reports 380
Jinzhu Yang

Construction and Experimental Verification of Automatic Classification
Process Based on K-Mer Frequency Statistics 391
Pengwei Zhu

A Strategy to Determine Priorities Among Multiple Goals: Approaches
from Network Models 401
Mucun Xie, Dachao Shang, and Shuyan Zeng

Innovative Application of Bayesian Algorithm in Network Security Risk
Assessment Model 412
*Haosheng Li, Qingqing Ren, Wei Chen, Yixuan Ma, Qingwang Zhang,
and Wanting Lv*

Power Fault Detection Method Based on Waveform Data and Expert System ... 425
Tengyue Gui, Weimin Xu, Husong Wang, and Haobin Xu

Transportation Network Scheduling System Based on Data Analysis 435
Shuting Xu

Program Structure Defect Localization and Repair Methods in Software
Security Reverse Analysis 444
Yan Li

Application of Data Mining in the Development and Management
of Software Engineering in Cloud Computing Platform 453
Qing Tan

Credit Rating Optimization Model Based on Deep Q-Network	464
<i>Yijiao Fan</i>	
Network Security Situation Automatic Prediction System Based on Artificial Intelligence	476
<i>Wenyue Qi</i>	
Research on Optimization Algorithm of Multi-agent System	485
<i>Jieru Wang</i>	
Time Series Decision Analysis Based on Linear Programming and SARIMA	496
<i>Shuai Li, Zeyuan Zhang, and Dongming Jiang</i>	
Artificial Intelligence-Driven Network Intrusion Detection and Response System	508
<i>Haokun Chen, Yiqun Wang, Shangyu Zhai, Wanrong Bai, Zhiqiang Diao, and Dongyang An</i>	
Identifying Consumer Behavior Patterns from Massive User Transaction Data Based on Data Mining Techniques	519
<i>Qi Wang</i>	
Hierarchical Scheduling Method of Power Emergency Based on Differential Evolution Algorithm	530
<i>Kuiwen Huang, Taiping Yuan, Haowen Yu, Jie Zhu, and Huajun Tang</i>	
Construction of Movie Knowledge Graph and Design of Recommendation System Based on Movielens Dataset Expansion	540
<i>Peng Dong</i>	
Design and Implementation of a High Concurrency Online Payment Platform Based on Distributed Microservice Architecture	551
<i>Tianyou Huang</i>	
Design and Implementation of a General Data Collection System Architecture Based on Relational Database Technology	561
<i>Yuxin Wang</i>	
System Design and Implementation of Particle Filter Algorithm Combined with Mean Shift in High-Precision Event Camera Positioning	573
<i>Shu Xu, Erlan Wang, and Haiming Zhang</i>	

Research on Optimization of Visual Space Fractal Design Algorithm
Based on Fractal Geometry and Complex Network Theory 585
 Huimin Chen, Zhenting Li, and Junlin Zhou

Binary Logistic Model of Smart Tourism Based on Data Information System ... 596
 *Danhong Chen, Lei Zhao, Yining Zhuang, Meilin Zhang, Yu Sun,
 and Xin Liu*

Author Index 607



Construction of a Comprehensive Safety Guarantee System for College Students Based on Digital Twin Technology

Junyan Song^(✉)

Shandong Business Institute, Shandong, China

57997092@qq.com

Abstract. Campus security incidents in colleges and universities are sudden and urgent in nature. Once they occur, they will have a great negative impact on the safety of college students and social stability. To address this problem, this paper proposes to build a comprehensive safety protection system for college students based on digital twin technology. The system first builds a digital twin model of the campus through 3D modeling technology, and then deploys IoT devices to collect video streams, entry and exit records, and emergency information on the campus. The collected data is transmitted to the big data center and efficiently stored and analyzed using cloud computing technology. On this basis, an intelligent early warning system is developed using support vector machines to predict and warn of potential security risks, while also building an emergency response system. The system also includes virtual safety education and training modules developed using digital twin technology, and regularly organizes emergency drills to test and optimize emergency response processes. Finally, through the suggestion box and online feedback system, we collect opinions and suggestions from students and faculty on the safety assurance system, and optimize and upgrade the safety assurance system based on the feedback and safety incident analysis results. After the comprehensive safety guarantee system for college students was established, the number of safety incidents decreased by about 52.2%, and the average student injury rate dropped from 1.735% to 0.54%. This paper confirms the actual contribution of the system in reducing safety incidents and lowering student injury rates, demonstrates the effective application of digital twin technology in campus safety management, and provides colleges and universities with a more intelligent and systematic safety assurance solution.

Keywords: Comprehensive Support System · Digital Twin Technology · Intelligent Early Warning · Emergency Response

1 Introduction

Higher education institutions should not only provide students with knowledge and skills, but also ensure their safety and health during their stay at school. As an emerging tool, digital twin technology provides a new solution for campus safety management.

This paper analyzes the application of digital twin technology in the field of campus safety, and also verifies the effectiveness of this technology in reducing safety incidents and reducing student injury rates through comparative experiments. This paper constructs and evaluates a comprehensive security assurance system based on digital twin technology, providing new ideas and tools for university safety management.

The paper is structured as follows: first, the research background and the current application status of digital twin technology in the field of campus safety are introduced; then the method of building a comprehensive safety protection system for college students and the overall architecture of the system are explained; and the experimental design is presented and the number of safety incidents and student injury rates before and after the construction of the protection system are compared; finally, the research findings are summarized and the potential and future development direction of digital twin technology in campus safety management are discussed.

2 Related Work

In today's rapidly developing society, safety education management has become an important issue in the field of education. Based on a brief overview of related work, Su [1] comprehensively analyzed the problems existing in safety education management and proposed corresponding innovative optimization measures based on actual conditions, in order to provide all-round protection for students' learning and life on the basis of improving the current work situation. Xie [2] conducted a comprehensive analysis of student safety education management from the perspective of new media, and combined existing problems and the characteristics of information dissemination on new media platforms to explore countermeasures for the efficient implementation of student safety education management. Xi and Cao [3] took safety as the starting point and deeply analyzed the safety hazards and common legal risks faced by college students in their daily life, study and social life, providing college students with a comprehensive and systematic set of safety knowledge and legal guidance. Based on the fact that some students have weak safety awareness, Fu and Cheng [4] found that public safety education can improve college students' skills in dealing with natural disasters, emergencies and man-made disasters, and enable them to have corresponding emergency response capabilities in emergency situations, thereby reducing casualties and property losses and ensuring the safety and stability of university campuses. Zhang et al. [5] conducted a survey and analysis on the current status of students' fire safety awareness and found that students currently have weak fire safety awareness, lack of fire safety knowledge and insufficient fire emergency response capabilities. They analyzed the hidden dangers and causes of fires in student dormitories and formulated corresponding fire prevention measures to improve students' fire safety awareness and the fire safety level of student dormitories and prevent fires.

Bhandal et al. [6] evaluated the application of digital twin technology in operations and supply chain management and identified the trends and value potential of this emerging research area. Khan et al. [7] provided a comprehensive overview of the concepts, classifications, challenges, and opportunities of digital twins for wireless systems. Aloqaily et al. [8] explored the integration of digital twins and advanced intelligent technologies to realize the metaverse. Mihai et al. [9] found that digital twin technology can

be used for single entities, end-to-end services, and multiple services, and can realize the analysis, design, and real-time monitoring and control of IoT services to achieve cost-effective and resource-optimized operations. Mendi et al. [10] explored the application of digital twin technology in the military field, especially in fault diagnosis and health monitoring of satellite systems. The above research not only covers the theoretical basis and practical strategies of safety education, but also demonstrates the potential of digital twin technology in improving the effectiveness of safety education and optimizing safety management. This study explores how to integrate digital twin technology into the safety education and management of college students. By building a comprehensive safety protection system for college students, it is expected to simulate and predict potential safety risks, provide safety education and emergency response training, and optimize the safety management process. This will not only enhance students' safety awareness and self-protection capabilities, but will also provide colleges and universities with a more intelligent and systematic security solution.

3 Methods

3.1 Construction of Digital Twin Model

Three-dimensional modeling uses computer programs to model physical systems in the real world into three-dimensional models. 3D modeling enables interactive experience of virtual scenes, provides visual and immersive interactive experience, and enables intuitive understanding and control of virtual scenes. In the comprehensive campus security system, 3D modeling technology creates a high-precision, high-fidelity model of the campus, including the campus's building structure, transportation system, and security facilities. Through three-dimensional modeling, the virtual-real integration of physical and information dimensions is achieved and the full reproduction of the operating system is completed. The digital campus building structure, transportation system and safety facilities are the basis for building a digital twin model [11]. Transforming the physical entity of the campus into a digital model, including the precise collection of indoor structural properties and graphic information of buildings. By using drone mapping and office software to process data, a detailed 3D model of the digital campus is constructed. These models include buildings, as well as transportation systems within the campus such as roads, sidewalks, and parking lots, as well as security facilities such as surveillance cameras, emergency alarm buttons, and access control systems. The digital process integrates data from campus security systems such as video surveillance systems, electronic patrol systems, and checkpoint systems to provide a picture of the campus' security situation monitoring, achieving comprehensive monitoring of "people, vehicles, land, events, and objects" across the entire campus.

3.2 IoT Device Deployment

The deployment of high-definition cameras is an important part of achieving real-time monitoring and intelligent early warning mechanisms. By installing high-definition video surveillance equipment in various teaching buildings, main roads and corners

of the campus, all-round monitoring of the campus can be achieved. These cameras not only provide real-time video streams, but also perform behavior recognition and anomaly detection through video analysis, thereby promptly identifying potential security issues. Deploying an intelligent access control system to strengthen access management on campus, ensure that only authorized personnel can enter specific areas, and provide safer and more convenient access control. The intelligent access control system is combined with the digital twin model to achieve real-time monitoring and analysis of the flow of people on campus, thereby improving the efficiency and effectiveness of campus security management. The emergency alarm button provides a quick response mechanism. In an emergency, pressing the emergency alarm button will immediately send out a signal for help. The button is equipped with a two-way voice intercom function, allowing on-site personnel to have real-time conversations with the alarm center. The emergency alarm button can also be combined with the digital twin model to achieve rapid positioning and response to emergency events and improve the campus' emergency response capabilities.

3.3 Comprehensive Safety Guarantee System for College Students

The comprehensive security protection system for college students is a comprehensive and systematic security management framework that deeply integrates digital twin technology, the Internet of Things, big data analysis, and cloud computing technologies. It includes intelligent warning systems, emergency response systems, virtual security education and training, and decision support systems, aiming to build an intelligent and responsive campus security environment. The system works collaboratively to prevent and respond to various campus security incidents and improve the safety level of students. Figure 1 shows the security system architecture:

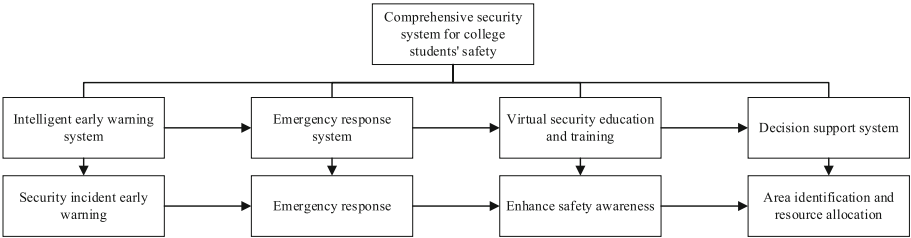


Fig. 1. The comprehensive security system architecture for college students

3.3.1 Development of Intelligent Early Warning System

Support vector machine (SVM) achieves classification by finding the maximum marginal boundary between different categories. In digital twin technology, SVM identifies patterns and abnormal behaviors from a large amount of collected data to predict and identify potential security risks. By analyzing campus surveillance video streams, SVM can identify illegal intrusions, abnormal gatherings and other behaviors and trigger warnings.

By analyzing data such as the flow of people on campus and abnormal behavior, the system identifies safety hazards and immediately sends warning notifications to the security management team to ensure that response measures are taken quickly. This early warning mechanism not only improves the response speed to campus security threats, but also intervenes before security incidents occur to reduce losses. In the SVM model, the identification of abnormal behavior can be further refined into the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} ||w||^2 + C \sum_{i=1}^n \xi_i \quad (1)$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, n \quad (2)$$

W is the normal vector of the hyperplane in SVM, b is the bias term in the SVM model, ξ_i is the slack variable for the i -th data point, which is used to handle data points that cannot strictly satisfy the hard margin condition, C is a regularization parameter that controls the trade-off between the margin width and the classification error, and n is the total number of data points in the dataset. x_i is the feature vector of the i -th data point, and y_i is the category label of the i -th data point.

3.3.2 Construction of Emergency Response System

The formulation of emergency handling processes and plans is the basis of the emergency response system. By integrating public safety data, analyzing and judging incidents, locating and alerting suspicious incidents, accurate deployment and online dispatch of police forces can be achieved. At the same time, relying on video fusion and feedback from the processing process, the time from discovering problems to solving them can be shortened, and a three-dimensional public safety prevention and control system based on the organic combination of terminal perception and three-dimensional scenes can be established. In addition, by upgrading the text plan to a visual form, it becomes a new type of emergency plan with the characteristics of operability, visualization, quantifiability, easy management, and easy sharing. When an accident occurs, multiple departments will coordinate their operations, quickly issue execution plans and provide special support with one click, so as to quickly locate the accident, complete the accident handling, and restore the campus to normal status.

Through digital twin technology, a series of virtual emergency drills were carried out, covering response strategies for various man-made emergencies such as terrorist attacks and school bullying [12, 13]. These drills enhance the ability of teachers and students to respond to emergencies. They also allow campus administrators to continuously optimize actual emergency response plans based on feedback from simulation results, striving to minimize casualties and property losses when real accidents occur. The digital twin platform uses a visual interface to display the location and status of various emergency resources in real time, queries the inventory of available emergency resources based on location data and sensor devices, provides underlying data support for emergency response plans for emergencies, and provides command, dispatch, and efficient resource allocation guarantees.

3.3.3 Virtual Safety Education and Training

By creating a virtual safety education environment and simulating various emergency situations, students can experience and learn how to deal with safety incidents without actual risks. This virtual education method improves students' safety awareness and self-protection ability. For example, laboratory safety emergency drills that combine VR virtual reality with actual operations can enhance laboratory safety emergency response capabilities. At the same time, schools use digital twin smart campus models to conduct virtual emergency drills, covering response strategies for various man-made emergencies such as terrorist attacks and school bullying [14]. These drills have improved the ability of teachers and students to respond to emergencies, and have also helped to continuously optimize actual emergency response plans, striving to minimize casualties and property losses when real accidents occur.

3.3.4 Decision Support System

The mean clustering algorithm analyzes the security incident data of different areas on campus, identifies high-risk areas, and provides decision support for resource optimization allocation. The digital twin smart campus platform analyzes the weak links in campus security management, identifies high-risk areas and time periods within the campus, and then guides the rational adjustment of security resource layout to ensure the accurate deployment and efficient operation of security forces [15]. At the same time, digital twin technology integrates and visualizes various data on campus. By building a digital twin model, various facilities and equipment on campus are digitally presented, and the real-time performance and alarm data of the equipment are displayed, achieving a panoramic perception of the campus security situation.

3.4 Data Collection and Processing

Data collection begins with various sensors and monitoring devices deployed on campus, which collect real-time information about the environment and infrastructure such as temperature, humidity, and energy consumption. These data are transmitted through switches, routers, firewalls and other network devices in the campus network as well as cloud service interfaces, as shown in Table 1.

Once the data is transferred to the big data center in the data transmission layer, the data processing and storage layer starts working. In this layer, the data is cleaned, transformed and aggregated, and the time series processing of the data supports real-time monitoring and historical data analysis. The processed data is stored in the database and used in subsequent analysis and query.

Cloud computing technology provides powerful data processing and analysis capabilities, updates the status of digital models in real time, and provides users with a visual display interface. Cloud computing technology supports efficient data storage and analysis, ensuring the real-time and accuracy of data processing. Cloud computing technology can be used to achieve real-time monitoring and early warning of campus safety, optimization of energy use and energy saving, and predictive maintenance of facilities and equipment.

Table 1. Campus monitoring data

Sequence Number	Collection Point	Sensor Type	Data Type	Sample Data
1	Academic Building A	Temperature Sensor	Temperature (°C)	23.5,23.7,23.6,... (Continuous Data)
2	Academic Building A	Humidity Sensor	Humidity (%)	55,56,55,... (Continuous Data)
3	Dormitory Building B	Energy Monitor	Electricity Consumption (kWh)	120, 125, 130,... (Hourly Accumulation)
4	Library	Temperature Sensor	Temperature (°C)	22.0,22.2, 22.1,... (Every 10 min)
5	Library	Humidity Sensor	Humidity (%)	60,59,60,... (Every 10 min)
6	Gymnasium	Light Intensity Sensor	Light Intensity (lx)	500, 520, 510,... (Every 30 min)
7	Canteen	CO ₂ Concentration Sensor	CO ₂ Concentration (ppm)	800,780,790,... (Every 15 min)

4 Results and Discussion

4.1 Warning Accuracy of Intelligent Warning System

In practical applications, the intelligent warning system can timely identify potential safety risks based on the data provided by the digital twin model. For example, when the population density in a certain area on campus exceeds a preset threshold, the system can immediately issue an early warning, prompting relevant departments to take evacuation measures to avoid safety accidents such as stampedes. In addition, the system also predicts and warns of equipment failures on campus, detects whether there are any abnormalities in the equipment in advance, and reduces safety accidents caused by equipment failures.

The intelligent early warning system uses the digital twin model to collect data such as personnel flow and environmental changes in these areas in real time. Figure 2 shows the accuracy of the early warning system in 20 different universities.

As shown in Fig. 2, the warning accuracy of the intelligent early warning system in different universities is above 95.3%, with the highest reaching 99.8%. These data results further confirm the key role of digital twin technology in the intelligent early warning system. In the experiment, the system is able to accurately capture and analyze data streams from various sensors and monitoring devices, quickly identify possible security threats, and issue timely warnings. When a fire occurs, the system analyzes data from smoke sensors and temperature sensors to accurately predict the location of the fire before it spreads.

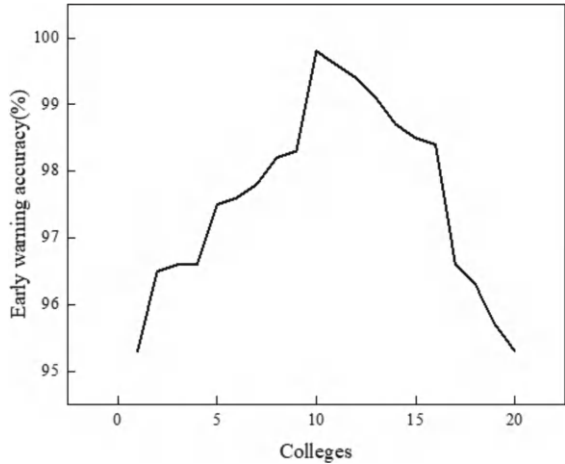


Fig. 2. Early warning accuracy

The intelligent early warning system not only performs well in technology, but also shows strong adaptability and flexibility in practical applications. The system can be customized according to the specific environment and needs of different universities to ensure that the accuracy of early warnings always remains at a high level. This high degree of customization enables the intelligent early warning system to be widely used in different types of university environments, providing college students with a safer learning and living environment.

4.2 Efficiency of the Emergency Response System

The emergency response system achieves all-round, real-time monitoring of the flow of people on campus, abnormal behavior, and potential security threats through the deep integration of high-definition video surveillance, face recognition technology, and behavioral analysis intelligent sensors. The system autonomously identifies security risks such as illegal intrusions and abnormal gatherings of students, and immediately sends early warning notifications to the security management team to ensure that response measures are taken quickly. It builds a full-chain and cross-departmental collaborative handling process, enhances the ability to coordinate responses to various accidents, disasters, natural disasters and comprehensive urban risks, and realizes the transformation of emergency management from “passive response” to “active prevention and control”. Figure 3 shows the average response time of emergency response systems in different universities.

Figure 3 shows that the average response time of the emergency response systems of these universities varies to a certain extent. The university with the shortest response time is 4.1 min, while the university with the longest response time is 9.4 min, which is a big difference. This reflects that there are great differences among different universities in the construction and operation of emergency response systems, including personnel quality, equipment configuration, technical level, and plan formulation. Secondly, judging from

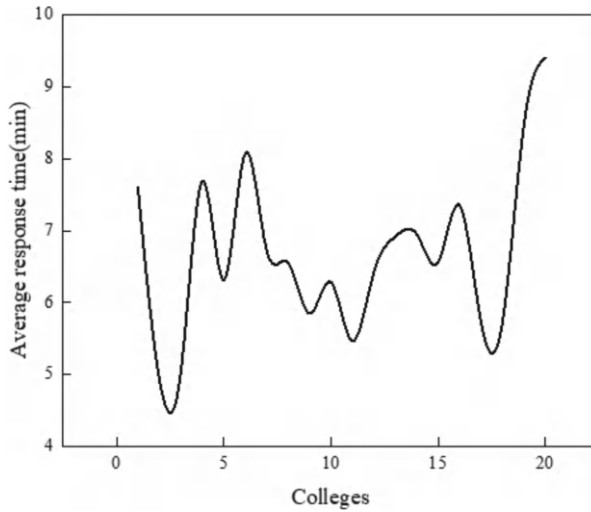


Fig. 3. Average response time

the distribution of the data, the average response time of emergency response systems in most universities is concentrated between 6 and 8 min, which shows that the emergency response systems of most universities can respond to emergency incidents in a relatively short period of time and have a certain emergency handling capability. However, some universities still have a long response time, exceeding 8 min, and some even reach more than 9 min, which will have an adverse impact on the timely handling of emergencies. For those colleges and universities with long response times, we need to explore the reasons behind them. Is it because of insufficient staff quality or backward equipment configuration? Is it because of limited technical level or incomplete emergency plan formulation? We need to find the root cause of the problem and take targeted measures to improve it.

4.3 Effect of Comprehensive Safety Assurance System

In order to quantify the application effect of the comprehensive safety protection system for college students, this paper conducts a comparative experiment to measure the changes in the number of campus safety incidents and the student injury rate before and after the construction of the protection system, so as to directly reflect the actual contribution of the comprehensive safety protection system for college students based on digital twin technology constructed in this paper to improving campus safety. The results are shown in Figs. 4 and 5 respectively:

From Fig. 4, we can see that among different universities, the number of security incidents has decreased significantly after the establishment of the security system. The total number of security incidents in the 20 universities before the construction is 435, while the total number of security incidents after the construction is only 208, a decrease of about 52.2%. Moreover, by analyzing the data of each university separately, it can be found that before the establishment of the college student safety protection system,

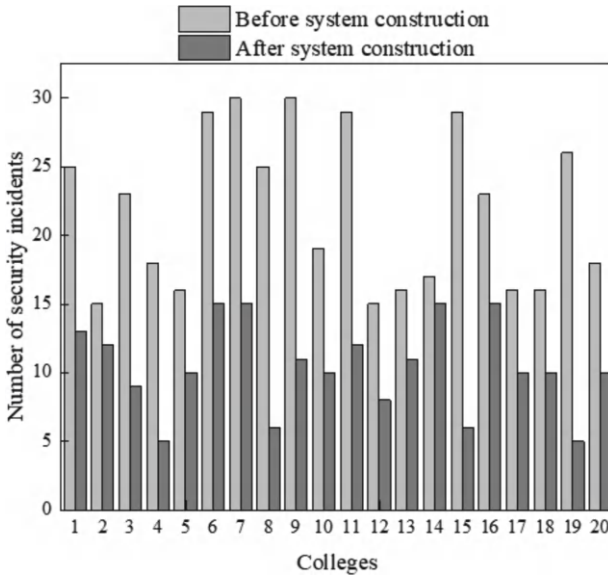


Fig. 4. Number of security incidents

college safety incidents occurred frequently. After the establishment of the protection system, the number of safety incidents in some universities has dropped significantly, and the decline has even exceeded the average level. This reflects that these universities have achieved remarkable results in building a security system, thanks to more perfect institutional design, more effective enforcement and more comprehensive security measures. At the same time, although the number of security incidents in some universities has decreased, the decline is relatively small. This means that these universities still have some shortcomings in the construction of the security system and may need to make more investments and improvements in system design, personnel training, equipment updates, etc.

From the overall trend of Fig. 5, after the construction of the college student safety guarantee system, the student injury rate of various colleges and universities has generally decreased. In the 20 sets of data before the construction, the average student injury rate is about 1.735%, while the average after the construction drops to about 0.54%, and the overall decline reaches about 1.195%. This data shows that the construction of a college student safety guarantee system is effective in reducing the student injury rate. The changes in the student injury rate before and after the construction show certain differences among different universities. Some colleges and universities have relatively high injury rates before the construction, but achieve a significant decline after the construction, such as the third college, which drops from 2.4% to 1%, a decline of more than 50%; although the injury rate of some universities has decreased after the construction, the decline is relatively small. Although the decline is not large, it also reflects the positive impact brought about by the construction of the security system. Based on the above, it is concluded that the construction of a safety protection system for

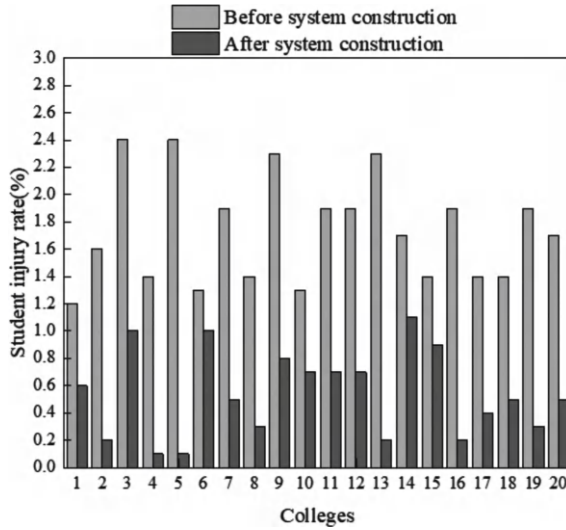


Fig. 5. Student injury rate

college students is effective in reducing the student injury rate, but there are differences in the changes in injury rates among different universities. In order to further improve the level of campus safety, more precise measures need to be taken according to the specific circumstances of different universities to ensure that every student can learn and grow in a safe environment.

5 Conclusion

This paper constructs and evaluates a comprehensive safety assurance system for college students based on digital twin technology, and ultimately solves the problem of how to use high-tech to improve the efficiency and effectiveness of campus safety management. This paper demonstrates a comprehensive, multi-level safety management system that predicts and responds to safety incidents on campus, reducing the number of safety incidents and student injury rates. The limitations of the paper are mainly reflected in the scale and scope of the experiment. Although the experiment covers a certain number of colleges and universities, it is limited by the region and sample size and cannot fully represent the actual situation of all colleges and universities. Future research can be conducted in a wider region and more colleges and universities, while considering more influencing factors to enhance the universality and applicability of the research.

References

1. Na, S.: Research on the strategies of college counselors to innovate and optimize the management of student safety education. *Indus. Technol. Forum* **23**(18), 245–247 (2024)
2. Jinzhi, X.: Research on student safety education management from the perspective of new media. *J. Jiamusi Vocat. Colle.* **40**(3), 142–144 (2024)

3. Zhen, X., Li, C.: Construction of college students' safety education system and legal awareness training strategy—A review of “Research on safety education and legal awareness training for college students”. *J. China Safe. Sci.* **34**(7), 251–252 (2024)
4. Chunlan, F., Lianhua, C.: Cultivation of public safety awareness among college students - Review of “Safety Education for College Students.” *J. Saf. Environ.* **24**(1), 411 (2024)
5. Hailong, Z., Ye Yanli, S., Jinfeng, H. Y.: Research on apartment fire prevention measures based on students' safety awareness. *Fire Protection (Electronic Edition)* **10**(2), 84–86 (2024)
6. Bhandal, R., Meriton, R., Kavanagh, R.E., et al.: The application of digital twin technology in operations and supply chain management: a bibliometric review. *Supply Chain Manage. Int. J.* **27**(2), 182–206 (2022)
7. Khan, L.U., Han, Z., Saad, W., et al.: Digital twin of wireless systems: overview, taxonomy, challenges, and opportunities. *IEEE Comm. Sur. Tutor.* **24**(4), 2230–2254 (2022)
8. Aloqaily, M., Bouachir, O., Karray, F., et al.: Integrating digital twin and advanced intelligent technologies to realize the metaverse. *IEEE Cons. Electr. Magaz.* **12**(6), 47–55 (2022)
9. Mihai, S., Yaqoob, M., Hung, D.V., et al.: Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Comm. Surv. Tutor.* **24**(4), 2255–2291 (2022)
10. Mendi, A.F., Erol, T., Doğan, D.: Digital twin in the military field. *IEEE Internet Comput.* **26**(5), 33–40 (2021)
11. San, O.: The digital twin revolution. *Natu. Computat. Sci.* **1**(5), 307–308 (2021)
12. Alcaraz, C., Lopez, J.: Digital twin: a comprehensive survey of security threats. *IEEE Comm. Surv. Tutor.* **24**(3), 1475–1503 (2022)
13. Liu, Y.K., Ong, S.K., Nee, A.Y.C.: State-of-the-art survey on digital twin implementations. *Advances in Manufacturing* **10**(1), 1–23 (2022)
14. Almasan, P., Ferriol-Galmés, M., Paillisse, J., et al.: Network digital twin: context, enabling technologies, and opportunities. *IEEE Commun. Mag.* **60**(11), 22–27 (2022)
15. Coorey, G., Figtree, G.A., Fletcher, D.F., et al.: The health digital twin: advancing precision cardiovascular medicine. *Nat. Rev. Cardiol.* **18**(12), 803–804 (2021)



Data Resource Value Evaluation Algorithm Based on Fuzzy Theory and Catastrophe Series Method

Lei Wang^(✉)

Chengdu Polytechnic, Chengdu, China
lerrywong@163.com

Abstract. The research on data resource value evaluation algorithms is of great significance for the construction of the current enterprise data resource accounting method system. The article proposes an evaluation algorithm for enterprise data resources, which is based on fuzzy theory and combines Catastrophe theory to avoid the dependence of traditional state evaluation algorithms on subjective weights, as well as the overly complex characteristics of Monte Carlo simulation and artificial neural network models. The algorithm can better reflect value changes from the perspective of state transitions. It comprehensively considers the factors that affect the value changes of data resources and can provide reasonable evaluations even in the absence of parameters. It is also simpler and easier to implement. Using actual data as an example, the effectiveness and feasibility of the algorithm were verified through the analysis of evaluation results and detection data.

Keywords: Fuzzy Theory · Catastrophe Level · State Evaluation · Data Resources

1 Introduction

Evaluating the value of data assets is of great significance for the development of enterprises. The importance of evaluating the value of data assets is reflected in many aspects. It not only relates to the strategic decision-making and resource allocation of enterprises, but also directly affects their market competitiveness, innovation ability, and economic benefits. Evaluating the value of data assets not only helps optimize resource allocation, enhance market competitiveness and innovation capabilities, achieve data monetization and economic benefits, but also guides data asset management and protection, and reveals potential value and business opportunities. Therefore, enterprises should attach great importance to data asset evaluation work, establish a sound evaluation system and methodology, and ensure the accuracy and reliability of evaluation results. On the basis of reviewing the research results of data asset evaluation methods, this article explores how fuzzy theory and mutation series method can be applied to the evaluation of data assets, and opens up new ideas for data asset evaluation.

2 Related Works

2.1 Development Status of Basic Concepts of Data Assets

The concept of data assets was first proposed by Richard E. Peterson (1974), who referred to them as assets with special attributes such as type, structure, etc., that can be digitized. In addition, he also proposed the characteristic of data as an asset, believing that it can circulate in the market like financial products such as bonds, thus possessing collateral value [1]. Gargano and Raggad (1999) argue that data can be an asset that not only creates economic benefits for businesses, but can also be exchanged [2]. Fisher (2009) proposed that data assets are assets that are not recorded in financial statements and can enhance a company's profitability [3]. Perrons and Jensen (2015) elaborated on the relationship between data resources and data assets, stating that valuable data resources are valuable assets [4].

Li Chunqiu and Li Ranhui (2020) focused on analyzing the five characteristics of data assets: non entity, reliance, diversity, processability, and value volatility. This is of great significance for selecting evaluation methods in the later stage to obtain reasonable evaluation results. When choosing an evaluation method, the characteristics of data assets should be considered in order to conduct a reasonable value assessment [5]. Gao Hua and Jiang Chaofan (2022) believe that the value of data assets is positively correlated with the cost value of data, while negatively correlated with the risk of data assets. In addition, the value of data assets is also affected by their frequency of use. Generally, the higher the frequency of use, the greater the value of data assets [6].

2.2 Development Status of Data Asset Evaluation

Longstaff and Schwartz (2001) first used the Least Squares Monte Carlo (LSM) method to evaluate the value of data assets, which to some extent solved the problem of excessive dependence on sales expense ratios and enhanced flexibility [7]. Si Yuxin (2019) first used the Analytic Hierarchy Process to separate the income brought by data assets and discount them to obtain the value of data assets [8]. Wang Xiaoxiao et al. (2019) first used artificial neural network models to evaluate the value of data assets. By constructing models for data value evaluation and fuzzy comprehensive evaluation using artificial neural networks, they improved the subjectivity issue in data asset value evaluation [9]. Zhao Chenyuan and Zhang Fulai (2023) used a combination of real option method and traditional data asset valuation method to calculate the value in two parts, considering the potential value of data assets and making the valuation more reasonable. The parameter connotation based on data asset related models is constantly developing and enriching [10].

2.3 Existing Data Asset Evaluation Algorithms

In the above-mentioned domestic and international research status, the research on data asset evaluation algorithms has a significant impact on the accuracy of data value evaluation. Scholars have proposed using the Analytic Hierarchy Process to break down the value of off balance sheet intangible asset groups [11]. However, due to its subjectivity,

some scholars later proposed Monte Carlo simulation and artificial neural network models to correct it, and then determined the division rate. However, these methods are too complex to calculate and lack practicality. Therefore, this article applies the Catastrophe series method to data asset evaluation to enrich the evaluation algorithm.

3 Methods

Catastrophe theory is an emerging mathematical branch that studies discontinuous phenomena, based on bifurcation theory, topology, and structural stability theory [12]. The research object of Catastrophe theory is the potential function $f(x, y)$ of a system, or the gradient system corresponding to the system. The commonly referred Catastrophe theory is actually an elementary Catastrophe theory, whose main mathematical origin is to classify critical points based on potential functions, and then study the characteristics of discontinuous states near various critical points, that is, a finite number of elementary Catastrophes. By combining the knowledge obtained in this way with theoretical analysis and observational data of discontinuous phenomena, mathematical models can be established to gain a deeper understanding of the mechanisms of discontinuous phenomena and make predictions [13]. The Catastrophe series method is a comprehensive evaluation method that decomposes the evaluation objectives into multi-level contradictions, and then combines Catastrophe theory with fuzzy mathematics to generate Catastrophe fuzzy membership functions. The system is quantified and recursively operated to obtain the Catastrophe membership function values that characterize the system's state characteristics, thereby ranking and analyzing the evaluation objectives.

3.1 Constructing a Catastrophe Evaluation Index System

According to the evaluation purpose, the overall evaluation indicators are subjected to multi-level contradiction decomposition, with the main factors at the forefront and the secondary factors at the back, arranged in a tree like hierarchical structure. From the overall evaluation indicators to the lower level indicators and then to the lower level sub indicators, only the lowest level indicators need to be known and used as raw data to obtain the final evaluation results using the Catastrophe series method. Decomposing indicators is to obtain more specific indicators for easier quantification. When the decomposition reaches a point where a certain sub indicator can be quantified, the decomposition can be stopped. Generally, the control variables for a state variable in a Catastrophe system should not exceed 4, and correspondingly, the decomposition of each level indicator should not exceed 4.

3.2 Determine the Catastrophe System Type of Catastrophe Evaluation Index System

The decomposed inverted tree structure can identify its function type based on the number of control variables. Thome proved in 1972 that there are seven basic Catastrophe systems, including folding Catastrophe system, apex Catastrophe system, swallowtail

Catastrophe system, butterfly Catastrophe system, parabolic umbilical point Catastrophe system, elliptical umbilical point Catastrophe system, and hyperbolic umbilical point Catastrophe system. Among them, there are four most common ones, namely folding Catastrophe system, apex Catastrophe system, swallowtail Catastrophe system, and butterfly Catastrophe system, among which the apex Catastrophe system is the most widely used. The models of the four Catastrophe systems, along with the number of control variables and corresponding normalization formulas, are shown in Table 1.

In Table 1, $f(x)$ is the potential function of the state variable x for one system; The coefficients a , b , c , and d of x are the control variables for the state variable; x_a , x_b , x_c , and x_d are the sudden state variable values corresponding to the four control variables a , b , c , and d , respectively. After determining the evaluation indicators, evaluators can determine the importance of each indicator based on experience. Among indicators at the same level, relatively important indicators are placed at the front and relatively secondary indicators are placed at the back. If one indicator can be fully reflected by one quantifiable indicator, then the system is considered a folding Catastrophe system; If one indicator is only decomposed into two sub indicators, the system is considered a Cusp Catastrophe system; If one indicator can be decomposed into three sub indicators, then the system is considered a swallowtail Catastrophe system; If one indicator can be decomposed into four sub indicators, then the system is considered a butterfly Catastrophe system. Indicator decomposition is used to obtain more specific sub indicators for quantification, and generally, the control variables for abrupt system state variables do not exceed four.

Table 1. Type of catastrophe system and its normalization formula

Type	Model	The number of variables	Normalization formula
Folding Catastrophe system	$f(x) = x^3 + ax$	1	$x_a = a^{\frac{1}{2}}$
Cusp Catastrophe system	$f(x) = x^4 + ax^2 + bx$	2	$x_a = a^{\frac{1}{2}}, x_b = b^{\frac{1}{3}}$
Swallowtail Catastrophe system	$f(x) = x^5 + ax^3 + bx^2 + cx$	3	$x_a = a^{\frac{1}{2}}, x_b = b^{\frac{1}{3}}, x_c = c^{\frac{1}{4}}$
Butterfly Catastrophe system	$f(x) = x^6 + ax^4 + bx^3 + cx^2 + dx$	4	$x_a = a^{\frac{1}{2}}, x_b = b^{\frac{1}{3}}, x_c = c^{\frac{1}{4}}, x_d = d^{\frac{1}{5}}$

3.3 Derive Normalized Calculation Formula from Bifurcation Equation of Catastrophe System

According to the theory of Catastrophe, for the thermal function of a Catastrophe system, all its critical points are combined into one surface. By making the first derivative 0

and $f'(x) = 0$, the equation of the equilibrium surface can be obtained, which is the critical point set of the Catastrophe function; By making the 2nd derivative 0 and $f''(x) = 0$, the singularity set of the equilibrium surface can be obtained. By solving two equations simultaneously and eliminating x , we can obtain a bifurcation point set equation represented by the state variables, which reflects the decomposition form of the relationship between the state variables and each control variable. The bifurcation point set equation indicates that when each control variable satisfies this equation, the system will undergo a sudden change. The normalized calculation formula can be obtained by decomposing the bifurcation point set equation. The normalization calculation formula is essentially a multidimensional fuzzy membership function, which can be used to perform quantitative recursive operations on the system to obtain the total Catastrophe membership function value that characterizes the system's state characteristics.

3.4 Use Normalized Calculation Formula for Evaluation

According to the theory of multi-objective fuzzy decision-making, in multi-objective situations, the strategy to meet the overall goal generally follows the principle of “taking the big from the small”; The principle of “taking the small from the large” should be used to calculate the values of various control variables for the same object using normalized calculation formulas; But for indicators with complementarity, their average is usually used instead; When comparing objects at the end, the principle of “taking the big from the small” should be used, that is, the evaluation objects should be sorted according to the score of the total evaluation indicators. From this, it can be seen that the determination of indicators at all levels is actually the result of a comprehensive ranking of the indicators at the next level.

4 Results and Discussion

4.1 Construction of Evaluation Index System for Data Resource Catastrophe

4.1.1 Important Information Reflecting the Status of Enterprise Data Resources

The important information that can reflect the status of enterprise data resources mainly includes customer relationships (A1), human capital (A2), and brand effects (A3). The customer relationship (A1) can be divided into main customer sales volume (B1), sales expense ratio (B2), and main business growth rate (B3); The main customer sales volume (B1) can be divided into the top 10 customer sales volume of the company (C1) and the total annual sales volume (C2), while the top 10 customer sales volume of the company (C1) can be divided into the top 5 customer sales volume of the company (D1) and the sales volume of the 6th to 10th customer sales volume of the company (D2); The growth rate of main business (B3) can be divided into the increase in main business revenue for the current period (D3), the increase in main business revenue for the previous period (D4), and the growth rate of user numbers (D5). The Catastrophe evaluation index system for enterprise data resources formed is shown in Fig. 1.

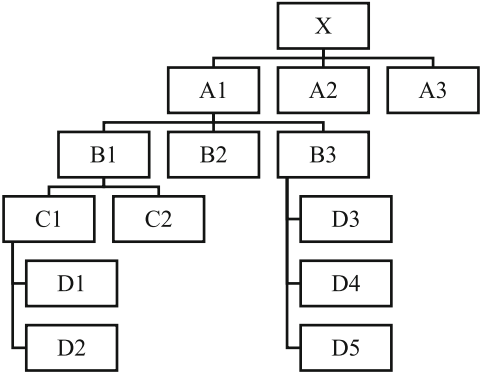


Fig. 1. Data resources evaluating indicator system based on catastrophe theory

4.1.2 Input Information for Catastrophe Evaluation Index System

From Fig. 1, it can be seen that the original input information in the Catastrophe evaluation index system includes D1, D2, C2, D3, D4, D5, B2, A2, and A3. Among them, human capital (A2) and brand effect (A3) are converted into unified data information through scoring methods. The main factors included in A2 and A3 are shown in Table 2.

Table 2. Factors of A2 and A3,subset

factor set	state factor
A2	Contribution value of labor force Proportion of employees with graduate education Employee training expense rate
A3	market share Per capita profit creation ratio of sales personnel

4.1.3 Determine Comment Set V

The purpose of fuzzy comprehensive evaluation is to obtain the best evaluation result from the set of comments based on comprehensive consideration of all influencing factors. After comprehensive consideration, the set of comments for evaluating the status of enterprise data resources is set as $V = \{v_1, v_2, v_3, v_4, v_5\}$, Among them, v_1, v_2, v_3, v_4 and v_5 correspond to the five evaluation levels of good, good, attention, abnormal, and severe, respectively. The corresponding maintenance strategies and suggestions are shown in Table 3.

Table 3. Maintenance strategy and suggestion corresponding to comment set

Comment	Status evaluation level	Strategy and Suggestions
V ₁	very good	The data resource status is normal and can be confirmed as an asset, entering the balance sheet accounting
V ₂	preferably	The status of data resources is basically normal and can be confirmed as assets, but it is still necessary to check the value realization of data resources
V ₃	be careful	There is uncertainty in the value realization of data resources, and evaluation should be strengthened to estimate the probability of value realization
V ₄	abnormal	There is significant uncertainty in the value realization of data resources, and the method of value assessment should be adjusted in a timely manner
V ₅	serious	Data resources are difficult to monetize and should not be included in the balance sheet accounting. Regular testing of their value is necessary

4.1.4 Membership Function

By substituting the original data and the transformed data volume information into the corresponding membership functions, the evaluation matrix of the corresponding evaluation indicators can be obtained. The membership function adopts a fuzzy distribution combining triangles and half trapezoids.

4.2 Determine the Type of Catastrophe System

From the Catastrophe evaluation index system in Fig. 1, it can be seen that, with complete input information, the types of Catastrophe systems include spike Catastrophe systems with control variables of (D1, D2) and (C1, C2); A swallowtail Catastrophe system with control variables of (D3, D4, D5), (B1, B2, B3), and (A1, A2, A3). If the input information is incomplete, determine the type of Catastrophe system based on the actual amount of information input. For example, in the absence of a sales expense ratio (B2), a swallowtail Catastrophe system with control variables (B1, B2, B3) will transform into a spike Catastrophe system with control variables (B1, B3).

4.3 Using Normalization Formulas for Recursive Quantization Operations

According to the corresponding Catastrophe type, use normalization calculation formula to transform the evaluation matrix corresponding to each evaluation index into a comprehensive evaluation matrix under the Catastrophe membership function. On this basis, following the principle of “complementarity”, the average value is taken to the Catastrophe membership function value of the upper level evaluation index, and so on. The total Catastrophe membership function value of the comprehensive evaluation result

is obtained by counting from the bottom to the top of the inverted tree structure. According to the “maximum membership degree principle” in fuzzy theory, the comment set element corresponding to the maximum membership degree is selected as the final result of the data resource in this state evaluation.

4.4 Validity Verification of the Algorithm in This Article

In order to verify the correctness of the algorithm in this article, according to the evaluation index system shown in Fig. 1, since there is no human capital (A2), brand effect (A3), sales expense rate (B2), main business growth rate (B3), and annual sales total (C2), the evaluation result of the data resource value evaluation status is represented as $X = A1 = B1 = C1$, which is a Cusp Catastrophe system with control variables (D1, D2). Using a data resource value evaluation example from a certain website, the feasibility and effectiveness of the algorithm in this article are demonstrated.

4.5 Evaluation Examples and Their Data Information

The input and storage information, internal analysis and conversion information, and external output information contained in the software, website (domain name), and internal company information owned by a certain website. Specifically, “input storage information” mainly includes data assets owned or controlled by the copyright of current affairs news, people’s information, cultural and artistic electronic works, etc. stored in its affiliated apps. They serve as the basis for the website’s profitability and cannot generate economic value for the website itself; Internal analysis conversion information “mainly includes the detailed user information data assets generated by the website after data analysis based on user click through rates and click preferences. Based on this, it attracts advertising advertisers to place advertisements. In addition, it also provides value-added services such as VIP membership, paid information, and online live streaming to meet user needs and achieve revenue; The ‘external output information’ mainly includes all data assets that have been filtered, modified, and improved by the website and presented to the platform directly connected to the user. It is based on ‘input storage information’ and is also a part that directly contributes to profits. It is the key to realizing the transformation from data to data assets.

4.6 Evaluation of Original Input Information and Quantification

Quantify human capital (A2) and brand effect (A3) through scoring methods. If the final score of each factor is set as Z , the cumulative score of each factor, i.e. the cumulative score of human capital, is T , and the cumulative score of the missing information factor, i.e. the proportion of missing information, is Q , then the scoring process and results are shown in Table 4.

According to Table 4, the human capital score is $T = 0.15$; The proportion of missing information is $Q = 0.6$; The total score of human capital based on the detection data, excluding the impact of missing data, is:

$$Z = \frac{T}{1 - Q} \times 100\% = \frac{0.15}{1 - 0.6} \times 100\% = 37.5\% \quad (1)$$

Table 4. The grading process

Serial Number	Proportion(%)	Score calculation process
1	15	$T = 0; Q = 0.15$
2	20	$T = 0; Q = 0.15$
3	10	$T = 0; Q = 0.25$
4	5	$T = 0; Q = 0.3$
5	7	$T = 0; Q = 0.37$
6	5	$T = 0; Q = 0.42$
7	8	$T = 0; Q = 0.5$
8	10	$T = 0; Q = 0.6$
9	10	$T = 0.05; Q = 0.6$
10	5	$T = 0.1; Q = 0.6$
11	5	$T = 0.15; Q = 0.6$

The completeness of information is $(1-Q) \times 100\% = 40\%$

The input data for the status evaluation of the data resources for this case is shown in Table 5.

Table 5. Input message of state evaluation method

Parameter	Value	Parameter	Value
D1	23.50	D5	59.1
D1	35.57	C2	33
D2	-99.16	B2	-
D3	25	A2	37.5
D4	50	A3	-

From Table 5, it can be seen that due to the missing information of control variables B2 and A3, the types of Catastrophe systems include spike Catastrophe systems with control variables (D1, D2), (C1, C2), (B1, B3), and (A1, A2); A swallowtail Catastrophe system with control variables (D3, D4, D5). The inverted tree structure of the evaluation model is shown in Fig. 2.

The calculation process of the evaluation algorithm is shown in Table 6. Substitute the input information of state evaluation into their respective membership functions to obtain the membership function matrix of each parameter. Determine the Catastrophe type based on the number of control variables and substitute it into the corresponding normalization calculation formula. Finally, obtain the Catastrophe membership function value of the corresponding intermediate control variable through the “complementary” mean.

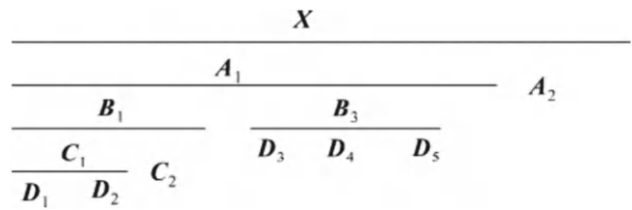


Fig. 2. Inverted tree structure of evaluation system

Table 6. Computing process of state evaluation based on catastrophe progression method

Parameter	Membership function vector	Substitute into the normalization calculation formula
D1	[0,0.5357, 0.6309,0,0]	(D1,D2 Cusp Catastrophe)
D2	[1, 0, 0, 0, 0]	[0,0.7319,0.7943,0,0] [1, 0, 0, 0, 0]
D3	[0,0.0625,0.6875, 0.6875, 0.0625]	(D3,D4,Ds swallowtail catastrophe)
D4	[0.5714,0.4000,0,0]	[0,0.2500,0.8292,0.8292,0.2500
D5	[0.4031,0.9709,0.3469,0,0]	0.8298,0.7368,0,0 0.7968,0.9926,0.7675,0,0]
C1	[05,0.3660,0.3971,0,0]	(C1,C2 Cusp Catastrophe)
C2	[0,0,0.1875,0.8125,0.5625]	[0.7070,0.6050,0.6301,0,0 0,0,0.5724 .0.9331,0.8255]
B1	[0.3535,0 3025,0.6013,0.4665,0.4127]	(B1,B2; Cusp Catastrophe)
B2	[0.5422,0 .6598,0 5322,0.2764,0.08331]	[0.5946,0.5500,0.7754,0.6830,0.6424 0.8154,0 .8706,0.8104,0 8304,0.4376]
A1	[0.7050,0.7103,0.7929,0.7567,0.5395]	(A1,A2 Cusp Catastrophe)
A2	[0,0,0,0,1]	[0.8396,0.8428,0.8904,0.8699,0.7345] [0,0,0,0,1]

In this article, the mean is used to achieve the effect of “complementarity”. For each operating state in the normalization calculation formula, we take:

$$X = \frac{\sum x_i}{n} \tag{2}$$

In the formula: $i = 1, 2, \dots, n$; n is the number of control variables; X_i is the membership function value of each control variable; X is the comprehensive evaluation membership function value. Calculate the Catastrophe level evaluation matrix of the next level evaluation index from the Catastrophe level evaluation matrix of the previous level evaluation index, and so on, until the top of the inverted tree structure, to obtain the total Catastrophe membership function value of the comprehensive evaluation result.

The total Catastrophe membership function value of the comprehensive evaluation result in this example is $X = [0.353, 0.355, 0.397, 0.435, 0.867]$. According to the principle of maximum membership degree, the maximum value is 0.8672, and the corresponding

comment set is “serious”. The correctness and effectiveness of the evaluation results were verified through the data resource status of the case, proving the practical feasibility of the evaluation algorithm.

5 Conclusion

This article proposes an evaluation algorithm for data resource status based on the Catastrophe series method. This evaluation algorithm fully considers the sales volume of major customers, human capital, and brand utilization, and is closely integrated with current value evaluation methods, with strong practicality. The evaluation algorithm based on the Catastrophe series method does not need to consider weights, only needs to stratify the evaluation indicators according to their impact on data resources and determine the master-slave relationship. By combining traditional fuzzy evaluation algorithms with Catastrophe series method, it avoids the dependence of traditional value evaluation algorithms on subjective weights and removes the differences caused by the subjectivity of traditional weights in the evaluation results. Compared to the simple fuzzy synthesis algorithm, the evaluation algorithm based on the Catastrophe series method is simpler and the credibility of the evaluation results is higher.

References

1. Peterson, R.E.: A cross section study of the demand for money: The United States, 1960–62. *The Journal of Finance* **29**(1), 73–88 (1974)
2. Gargano, M.L., Raggad, B.G.: Data mining-a powerful information creating tool. *Oclc Systems & Services* **15**(2), 81–90 (1999)
3. Fisher, T.: *The data asset: how smart companies govern their data for business success*. Wiley, New York (2009)
4. Perrons, R.K., Jensen, J.W.: Data as an asset: What the oil and gas sector can learn from other industries about “Big Data.” *Energy Policy* **81**, 117–121 (2015)
5. Li, C., Li, R.: Research on data asset value evaluation based on business plan and revenue: a case study of data asset value evaluation of a unicorn company. *China Asset Appraisal* (10), 18–23 (2020)
6. Gao, H., Jiang, C.: Valuation of Data Assets from the Perspective of Application Scenarios. *Finance and Accounting Monthly* (17), 99–104 (2022)
7. Longstaff, F.A., Schwartz, E.S.: Valuing American options by simulation: a simple least-squares approach. *Review of Financial Studies* (2001)
8. Si, Y.: Research on the value evaluation model of data assets in internet enterprises. Capital University of Economics and Business (2019)
9. Xiaoxiao, W., Hongjun, H., Shuchen, Z., Jing, W.: Research on value evaluation of big data based on fuzzy neural network. *Sci. Manage.* **21**(02), 1–9 (2019)
10. Chenyuan, Z., Fulai, Z.: Internet enterprise data asset value evaluation based on binary tree option pricing model. *China Asset Evaluation* **09**, 51–60 (2023)
11. China Asset Appraisal Association Guiding Opinions on Data Asset Evaluation. *Zhong-pingxie* [2023] No. 17
12. Zhang, J.R., Yang, S.J., Zhou, L.: The internal logic and system reconstruction of the evolution of university patent system in China. *Sci. Technol. Progress and Policy* **40**(21), 99–107 (2023)
13. Deng, S.L., Wang, F., Wang, H.W.: Identification methods of artificial intelligence generated content in online communities. *Documentation, Information & Knowledge* **41**(2), 28–38 (2024)



Comparative Study of ARIMA Model and Long Short Term Memory Network (LSTM) in Economic Management

Xiwen Wang^(✉)

Liaoning Vocational University of Technology, Jinzhou 121007, Liaoning, China
56010410@qq.com

Abstract. This paper collects and organizes economic data and establishes ARIMA (Auto Regressive Integrated Moving Average) model and LSTM (Long Short-Term Memory) model to compare their accuracy and stability in predicting the future trend of economic indicators. The results of the study indicate that in some cases, the LSTM model more accurately captures the long-term dependencies in the time series data and improves the forecasting results, while in some cases, the ARIMA model performs more stably and reliably. The study concludes that the LSTM model predicts the results of the indicators of economic management with an accuracy of up to 96%, but it also has shortcomings such as higher complexity. Therefore, choosing the appropriate model depends on the specific data characteristics and forecasting needs, and it is recommended to consider the advantages and disadvantages of ARIMA model and LSTM model in practical applications, and choose the most suitable model for the forecasting work in the field of economic management. This paper provides a more scientific forecasting method and decision-making reference for economic management decision-making.

Keywords: Economic Management · Long Short-Term Memory · Auto Regressive Integrated Moving Average

1 Introduction

With the rapid development of data science and artificial intelligence technology, various kinds of predictive models are gradually introduced in the field of economic management to help decision makers make accurate predictions and analysis. In this paper, a comparative study of two commonly used time series forecasting models, ARIMA model and Long Short-Term Memory Network (LSTM), will be conducted with the aim of exploring the effects, advantages and disadvantages of their applications in economic management. The main contribution of this paper is to provide a more scientific and accurate prediction method and decision basis for economic management decisions by comparing the performance of ARIMA model and LSTM model in terms of prediction accuracy, stability and training time.

The organization of this paper is as follows: this paper firstly introduces the background and significance of the research, analyzes the problems in traditional research, as well as the analysis basis and research motivation of this paper. This paper elaborates the theoretical basis and modeling methods of ARIMA model and LSTM model, as well as the application scenarios and advantages and disadvantages in economic management. This paper describes the operational steps in the data preparation and preprocessing stage, including data sources, processing methods and feature selection. This paper shows the comparative results of ARIMA model and LSTM model in the prediction of economic indicators, including the comparative analysis of prediction accuracy, training time and model complexity. The paper gives a conclusion and an outlook, summarizing the main findings and revelations of the paper, as well as the direction and focus of future research. This paper addresses the problem of how to choose appropriate forecasting models to improve the forecasting accuracy and stability of economic indicators in the field of economic management. The innovation of this paper is to compare and analyze the application of ARIMA model and LSTM model in economic management, and verify the performance of the two models in predicting economic indicators through empirical analysis and comparison. The technical solution includes preprocessing and feature selection of official economic data, modeling and forecasting using ARIMA model and LSTM model, and finally comparing the advantages and disadvantages of the two models to provide decision makers with more accurate forecasting methods and decision support. Through this paper, we hope to provide more scientific and effective methods and technical support for forecasting and analyzing in the field of economic management.

2 Related Work

In the recent years, there are many scholars who have studied economic management. Diah AM explored in depth how to promote economic management in developing countries [1]. Sustiyatik E explored the dynamics of economic management in non-formal education with a resource management analysis based on sustainable development [2]. Aliyeva M proposed the concept and theoretical analysis of the scientific and technological development management system of industrial enterprises [3]. He H compared the application cases of blockchain in different economic management fields, listing the challenges and limitations, aiming to comprehensively explore the challenges it faced [4]. Chai H put forward the importance of agricultural economic management, analyzed the development status of agricultural economic management and the characteristics of agricultural economic management under different development stages [5]. The above studies lacked in-depth analysis and verification of actual cases, which led to the limitations and lack of practicality of the research conclusions.

There are also a number of experts who have discussed both ARIMA and LSTM techniques. Nkongolo M used ARIMA model for growth prediction of user data usage [6]. Fauzani S P used ARIMA method to forecast the price of rubber producers in Riau Province in 2023 [7]. Albeladi K used LSTM and ARIMA for time series forecasting and found that LSTM was suitable for dealing with complex time series data and was able to capture long term dependencies, while ARIMA was suitable for smoother data

[8]. In order to eliminate subjectivity in predicting future crude steel, steel, and pig iron production, Chen H objectively predicted future related production based on the ARIMA model in time series [9]. Chen HF established LSTM prediction model for high frequency subsequence and ARIMA prediction model for low frequency subsequence [10]. The above discussion lacks to explore the rationality and accuracy of model selection, which affects the accuracy and reliability of prediction results.

3 Method

3.1 Data Preparation

In order to ensure the authenticity and validity of the data, this paper is based on the four official economic statistics mentioned above. Since there are no other variables in the sample, we do not take into account the effect of price elements on economy indexes. But in reality, the main effect is the price and the exchange rate. Therefore, we choose the price and the exchange rate as the research target. In order to enhance the precision of the study, we first carry out the following steps: First, we normalize the 4 official economy data with LS, eliminate the obvious non-relevant variables, then smooth them with standard time sequence data, and then carry on the training and prediction with LSTM model. In order to test the validity of this model in economics, we have carried out single-variable and multi-variable analysis on the above-mentioned 4 official economic data.

In order to eliminate the interactions between economic indicators and reduce the multicollinearity between the dependent and independent variables, the sample data were first standardized using the least squares method (Least Square). Due to the correlation of the sample data itself, the data processed through the least squares method no longer have linear correlation. In order to further eliminate multicollinearity among variables, the standardized time series data were smoothed. In smoothing the economic indicators, Eviews 8.0 software was chosen to be used as follows:

$$\begin{aligned} X_t &= c + \delta_p y_{t-p} + \theta_q \varepsilon_{t-q} \\ i_t &= \sigma (W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) f_t \end{aligned} \quad (1)$$

where δ is the coefficient of the autoregressive term, θ is the coefficient of the moving average term, X_t is the time series data, i_t is the input gate, W is the weight parameter, and b is the bias term. In this paper, the correlation coefficient matrix of the sequence is calculated first, and then determine whether the sequence is smooth or not. When calculating the correlation coefficient matrix, in order to ensure the accuracy of the calculation results, the first unit root test is performed, if the test result is the significance level $\alpha = 0.05$, then it means that there is a unit root in the sequence, and further smoothing can be carried out. The correlation coefficient matrix is first tested for unit root using the ADF (Augmented Dickey-Fuller) test. If the test result is at the significance level $= 0.05$, then the series is a smooth series; otherwise, the series is considered to be unstable. For the smoothed time series data, autocorrelation and partial autocorrelation tests were first performed by Eviews 8.0 software.

The standardization of economic data refers to the processing of time-series data so that they satisfy a normal distribution, i.e., the mean and standard deviation of the data are the value of the original indicator and the standard deviation of that indicator, respectively. For the four types of official economic data mentioned above, they are first standardized using the least squares method to obtain standardized data. There is no great difference in the standardized data, indicating that the standardized time series data satisfy the normal distribution. Therefore, the LSTM model is next used to model and analyze them. However, since the above four official economic data are all first-order single-integer series, they need to be differentiated to make them first-order difference series. In this process, a unit root test is first performed on the original time series. Since the time series does not have a unit root, it is analyzed using the basic principles of the unit root test. There is a lagged variable in the original time series which is not directly related to the feed-forward variable. Therefore the ADF test is used to determine whether the time series has a smooth nature. When it does not satisfy the smooth nature, it needs to be smoothed [11].

In order to further verify whether each economic indicator is smooth or not, this paper firstly conducts the unit root test for each variable. When building the LSTM prediction model, the first-order difference time series data of these economic indicators cannot be used directly for training and prediction. In this paper, the LSTM prediction is carried out after the differential smoothing of each economic indicator. In this paper, after univariate and multivariate analysis of the above four official economic data, the above data are trained and predicted using LSTM model. The training process of LSTM model is divided into three parts: input feature data, initialization of implied nodes, and carrying out training. After the training of LSTM model is completed, the historical economic data are predicted by LSTM model and compared with the original data. In the following, the prediction results of the LSTM model are compared and analyzed with the original data.

3.2 ARIMA Modeling

ARIMA model is one of the most commonly used models for modeling and forecasting non-stationary time series data, which is usually used to fit time series data with a smooth trend, such as interest rates, bond yields, stock index returns, etc., and is also used to forecast time series data with a volatile trend. The ARIMA model obtains the parameters by fitting the original data and then estimates the parameters to predict the future trend. The basic idea of the model is to smooth the time series first, and then model and forecast the smoothed time series to get the prediction results, the ARIMA model is shown in Fig. 1.

The comparison of the fitting results shows that the ARIMA(1,1) model has some advantages in fitting and forecasting compared to several other commonly used forecasting models. If the white noise test is used to examine whether it is reasonable to use the white noise test when fitting the ARIMA(1) model, it will be found that: regardless of whether the original data are autocorrelated or not, whether they are smooth time series or not, whether they are white noise or not, and which test is used to test the test, etc., the ARIMA(1) model and several other commonly used forecasting models have shown certain advantages [12, 13].

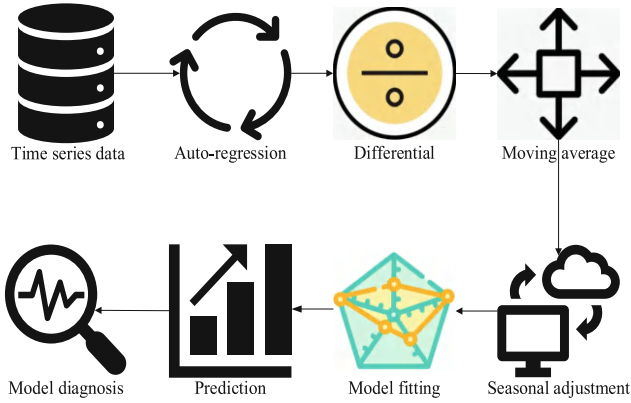


Fig. 1. ARIMA model

3.3 Modeling of LSTM Model

The LSTM model contains a special “forgetting gate”, when the output value is close to the historical value, the LSTM model will store more information, i.e., forget less information; on the contrary, when the output value is more different from the historical value, the LSTM model will store less information, i.e., forget more information. This mechanism better captures the long term dependencies in the time series. In LSTM model, data is the most important part in the input layer. The inputs to the LSTM model are generated from historical data. Historical data is computed by the forgetting gates in the LSTM model. This paper uses a number of statistical methods to obtain historical data, including regression analysis, trend analysis, and difference in means. Since these statistical methods usually require great labor and time costs, they are usually used as preprocessing of historical data, and LSTM models are used to predict these preprocessed historical data. After normalizing these historical data, some simple statistical methods are then used to extract the long-term dependencies in the time series [14].

In time series modeling, ARIMA (p,d,q) model is usually used. LSTM is mainly applied to forecasting of time series. In LSTM, there are two types of connections between the input layer and the hidden layer: a unidirectional connection, where the output layer can only interact with the previous layer of inputs, and a bidirectional connection, where the output layer interacts with both the previous and the subsequent layers. Each neuron of LSTM contains three gates to process input data and time series data respectively. In LSTM, a special “forgetting gate” is used to handle the long-term dependency between the input data and the time series data. The “Oblivion Gate” compares the output values of the input data with the historical values in the time series data. The training process of the LSTM model is a preprocessing phase, the main purpose of which is to remove the noise from the data, and also to perform the necessary cleaning and feature extraction of the data. In this stage, data cleaning is realized by back propagation algorithm and also model training is realized by Adam algorithm. Finally, the validity of the model is verified by a series of prediction results. The LSTM model has a good prediction ability and high prediction accuracy. In this series, the short-term volatility shows a certain degree of increase, while the long-term volatility decreases.

This indicates that the LSTM model captures the long-term dependence better and also the short-term volatility trend, and thus is informative for economic forecasting [15].

4 Experimental Results and Discussion

4.1 Accuracy Analysis

In this paper, by organizing and filtering the collected economic data, the monthly data including GDP, value added of industry, investment in fixed assets, and total retail sales of consumer goods are selected as the experimental objects, and the ARIMA and LSTM models are established to predict the future trend of each index. Therefore, this paper mainly evaluates the effectiveness of ARIMA and LSTM models in predicting the future trend of economic indicators in terms of the accuracy and stability of the prediction results, as shown in Figs. 2 and 3.

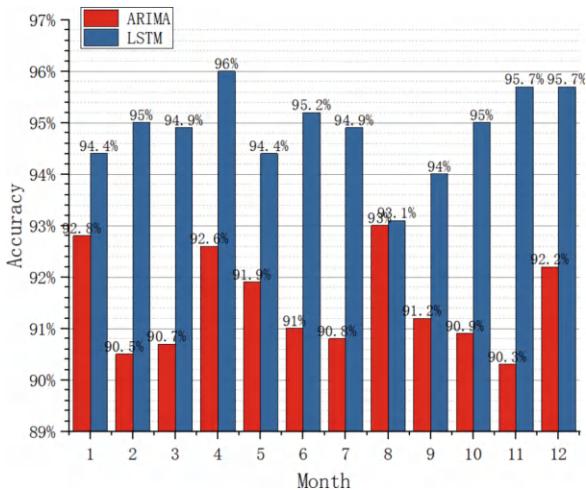


Fig. 2. Accuracy of prediction results

In Fig. 2, the accuracy of the prediction results of the ARIMA model for the indicators of economic management is the highest 93%, followed by 92.8%, and the lowest 90.3%, and the calculated average accuracy is 91.49%; the accuracy of the prediction results of the LSTM model for the indicators of economic management is the highest 96%, followed by 95.7%, and the lowest 93.1%, and the calculated average accuracy is 94.86%.

4.2 Stability Analysis

In Fig. 3, the stability of the prediction results of the ARIMA model for the indicators of economic management is the highest 94.9%, followed by 94.5%, and the lowest 93%, and the calculated average stability is 93.9%; the stability of the prediction results of the LSTM model for the indicators of economic management is the highest 97.8%, followed by 97.6%, and the lowest 95.2%, and the calculated average stability is 96.6%.

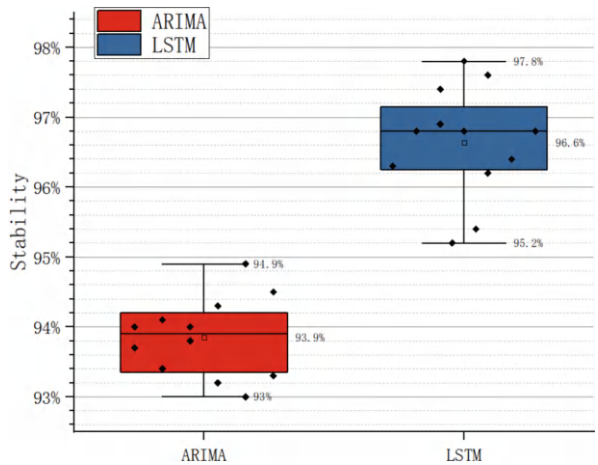


Fig. 3. Stability of prediction results

4.3 Comparison of Economic Management Model Performance

Finally, this paper compares the various performances of the traditional model, ARIMA model and LSTM model as shown in Table 1.

Table 1. Comparison of economic management model performance

Feature	Traditional	ARIMA model	LSTM
Training time	5 min	2 min	1 h
Number of parameters	10	40	1000
Model complexity	Low	Low	High
Feature importance	6.5	8.7	9.6
Model interpretability	7.3	8.8	9.3

In Table 1, the ARIMA model shows the advantage of being applicable to simple scenarios with shorter training time and lower model complexity, and is better than traditional methods in terms of feature importance and model interpretations, while the LSTM model excels in terms of feature importance and model interpretations, but with long training time, large number of parameters, and high model complexity, but these features make it more applicable to complex scenarios. Different factors should be taken into account when selecting a model in order to choose the most suitable model for forecasting and analyzing work in economic management.

4.4 Experimental Discussion

The results of this study show that the LSTM model is superior to the ARIMA model in terms of prediction accuracy and stability, with the highest accuracy rates of 96% and

93%, and the stability rates of 97.8% and 94.9%, respectively. This difference stems from the fact that the LSTM model can effectively capture long-term dependencies and complex nonlinear characteristics in time series, while the ARIMA model performs better in processing stationary time series and is suitable for simple scenarios. Therefore, when faced with the diversity and complexity of economic data, LSTM has more advantages, but its training time and model complexity are also relatively high.

This study provides a more scientific prediction method for the field of economic management, emphasizing that the data characteristics and specific application scenarios should be considered comprehensively when selecting a prediction model, which has an important impact on improving the effectiveness of economic decision-making. By deeply comparing the advantages and disadvantages of the ARIMA and LSTM models, the study provides decision makers with more accurate prediction tools to help them make more informed choices in a dynamic and complex economic environment. This paper improves the understanding of economic data and promotes in-depth analysis of complex economic phenomena, making the decision-making process more data-based and empirical. Future research will explore more advanced deep learning models and other time series analysis methods based on the results of this study to further improve the accuracy and practicality of economic forecasts. At the same time, it will help improve the theoretical framework of economic management and provide more forward-looking decision-making support for policymakers, ultimately promoting innovation and progress in economic development and management.

5 Conclusion

Combining the above experimental results and comparative analysis, it is learned that there are some differences between ARIMA model and LSTM model in terms of prediction effectiveness and stability in economic management. Through this paper, it is found that the LSTM model performs better in terms of prediction accuracy and stability, especially in the capture and prediction of long-term dependence relationships with obvious advantages. In contrast, the ARIMA model is closer to traditional methods in terms of training time, model complexity, feature importance and interpretability, and is suitable for forecasting work in simple scenarios. Considering the characteristics, advantages and disadvantages of the two models, it is crucial to choose the model suitable for specific needs for economic management forecasting. In the future research and application, the parameter settings of ARIMA and LSTM models will be further optimized to improve the prediction accuracy and stability of the models; at the same time, more economic factors and external variables will be combined to improve the prediction ability of the models. In addition, the application of other deep learning models and time series analysis methods are explored to enhance the forecasting effectiveness and decision support capabilities in the field of economic management. Through continuous research and practice, data science and technology can be better utilized to provide more scientific and accurate prediction and analysis methods for economic management decision-making, and to promote the progress of economic development and management. Thanks for the support and participation in this paper, and we look forward to sharing and exchanging more in-depth research and practice results in the future.

References

1. Diah, A.M., Away, J.L., Kadang, T.: Economic management in developing economies: strategies for sustainable growth and development. *Join: J. Soc. Sci.* **1**(4), 363–372 (2024)
2. Sustiyatik, E., Jauhari, T., Gupta, S.: Dynamics of economic management in the context of non-formal education: an analysis of resource management for the sustainability of education programs. *J. Nonformal Educ.* **9**(2), 207–216 (2023)
3. Aliyeva, M.: The role of econometrical modeling in increasing the efficiency of the economic management mechanism of innovation processes in industrial enterprises. *Eurasian J. Technol. Innov.* **1**(11), 84–89 (2023)
4. He, H.: The application of blockchain technology in economic management innovation. *Sci. Technol. Innov. Product.* **45**(2), 87–89 (2024)
5. Chai, H., Li, K., Yang, L., et al.: Research on the development trend of agricultural economic management under the background of rural revitalization. *J. Beijing Vocat. Colle. Fin. Trade* **40**(1), 16–21 (2024)
6. Nkongolo, M.: Using arima to predict the growth in the subscriber data usage. *Eng.* **4**(1), 92–120 (2023)
7. Fauzani, S.P., Rahmi, D.: Penerapan metode ARIMA dalam peramalan harga produksi karet di provinsi riau. *Jurnal Teknologi dan Manajemen Industri Terapan* **2**(4), 269–277 (2023)
8. Albeladi, K., Zafar, B., Mueen, A.: Time series forecasting using LSTM and ARIMA. *Int. J. Adv. Comput. Sci. Appl.* **14**(1), 313–320 (2023)
9. Chen, H., Hu, J., Wang, S., et al.: Research on CO₂ Emission characteristics and emission reduction path of China's steel industry — Based on ARIMA-LEAP model. *Environ. Sci. China* **44**(6), 3531–3543 (2024)
10. Chen, H.F., Wang, H., Li, Y., et al.: A short-term wind speed prediction study with a combined two-step decomposition and ARIMA-LSTM. *J. Solar Ener.* **45**(2), 164–171 (2024)
11. Kleiner, G.B.: System paradigm as a theoretical basis for strategic economic management in modern conditions. *Management Sciences* **13**(1), 6–19 (2023)
12. Sharma, A.K., Punj, P., Kumar, N., et al.: Lifetime prediction of a hydraulic pump using ARIMA model. *Arab. J. Sci. Eng.* **49**(2), 1713–1725 (2024)
13. Mgammal, M.H., Al-Matari, E.M., Alruwaili, T.F.: Value-added-tax rate increases: a comparative study using difference-in-difference with an ARIMA modeling approach. *Humanit. Soc. Sci. Comm.* **10**(1), 1–17 (2023)
14. Torres, J.F., Martínez-Álvarez, F., Troncoso, A.: A deep LSTM network for the Spanish electricity consumption forecasting. *Neural Comput. Appl.* **34**(13), 10533–10545 (2022)
15. Rostamian, A., O'Hara, J.G.: Event prediction within directional change framework using a CNN-LSTM model. *Neural Comput. Appl.* **34**(20), 17193–17205 (2022)



The Application and Empirical Study of Causality in the Theory Graph

Guijiao He^(✉)

Software Engineering Institute of Guangzhou, Guangzhou, Guangdong, China
hgjpaper@163.com

Abstract. The purpose of this paper is to study the application of causality in matter-of-fact mapping, and improve the inadequacy of traditional mapping that relies only on event co-occurrence relations. By introducing the causal inference method, we construct a more accurate causal matter-of-fact mapping and validate it in practical applications. The experimental results show that the deep learning model based on BERT (Bidirectional Encoder Representations from Transformers) achieves an F1 value of 0.875 in causal extraction, which outperforms CNN (Convolutional Neural Network)'s 0.825 and RNN's 0.825. In terms of graph optimization, PageRank optimization increases node importance from 0.05 to 0.075, HITS (Hyperlink-Induced Topic Search) optimization increases graph connectivity from 0.70 to 0.85, and the average path length is shortened from 3.5 to 3.0. It can be seen from the data conclusions that optimized causal graph is more effective and precise in representing event relationships and logical structures.

Keywords: Causality · Matter Diagram · BERT Algorithm · Pagerank Optimization

1 Introduction

With the advent of the big data era, the explosive growth of information makes it especially important to mine valuable information from massive data. As an important tool for revealing the intrinsic connection between things, causality plays a key role in research in various fields. As a structured knowledge base that describes events and the relationships between them, the causal map provides a powerful support for understanding and analyzing complex events. Studying the application of causality in matter-of-fact mapping not only helps to improve the accuracy and usefulness of the mapping, but also provides a reliable basis for tasks such as decision support and knowledge reasoning.

In this paper, we propose methods that combine deep learning and graph theory optimization by deeply exploring the application of causality in matter-of-fact graphs. Through empirical studies, we validate the effectiveness of these methods in causality extraction and graph construction. In particular, the significant performance of BERT model in causality extraction and the practical effect of PageRank and HITS algorithms in graph optimization are utilized to further enhance the capability of matter-of-fact graphs in event relation representation and logical structure analysis.

The structure of this paper is as follows: first, this paper introduces the basic concepts of causal and matter-of-fact mapping and their applications in existing research. Then, this paper describes in detail our research methodology, including data collection, preprocessing, causality extraction, graph construction and optimization. Then, the experimental results are presented and analyzed to verify the validity of the proposed method. Finally, this paper discusses the problems in the research and future improvement directions, and summarizes the main contributions of this paper.

2 Related Works

In recent years, the research on matter-of-fact mapping mainly focuses on the identification and utilization of event co-occurrence relationships. For example, traditional recommender systems estimate users' ratings of items based on observed ratings from the crowd, but hidden confounding factors may lead to systematic bias. Zhu Y et al. suggested introducing causal reasoning to address the effect of unobserved confounding factors [1]. The existing visual question answering methods are often affected by cross modal false correlations and overly simplified event level reasoning processes, which cannot capture the temporal, causal, and dynamic nature of events in videos. To address these issues, Liu Y and Li G proposed a cross modal causal inference framework for handling event level visual question answering tasks [2]. Liu Y and Wei Y S et al. comprehensively reviewed existing visual representation learning causal inference methods, including basic theories, models, and datasets, while discussing the limitations of these methods and datasets [3]. Zhang Shiying proposed a support path that combines theoretical knowledge graph and network public opinion analysis to address the problems of static knowledge, fuzzy reasoning, and spatial singularity in current emergency intelligence support[4]. An Biao proposed a method for constructing a trend graph of bulk agricultural product prices, using news data from the pig market as an example, to address the issue of insufficient analysis of the impact of news events in the study of price fluctuations of bulk agricultural products[5]. Deng Z et al. introduced Compass, a novel visualization analysis method for in-depth analysis of dynamic causality in urban time series [6]. Chen Yue et al. focused on solving high school geography causal short answer questions. This task requires knowledge integration and multi hop causal reasoning, and they have defined an abstract causal graph to represent causality. By pre training language models, they automatically extracted abstract reasoning graphs suitable for high school geography causal short answer questions from the corpus, achieving multi-source knowledge integration [7]. Yang J et al. believe that causality, as an important component of human cognition, often appear in texts. Organizing these causality from texts can help construct causal networks for predictive tasks [8]. However, there are still shortcomings in the existing research on causality recognition and graph construction methods.

The literature indicates that causal reasoning methods have potential in solving complex event relationship recognition problems. For example, Yang Z et al.'s study used the Granger causality test in quantile method to investigate the determinants of carbon dioxide emissions in China. The results showed that urbanization, financial development, and trade openness are the main determining factors of China's carbon dioxide emissions [9]. Wesiah S used quarterly data from the first quarter of 1963 to the first

quarter of 2015 to study the causality between financial development and economic growth in the UK. By using Johansen cointegration test and Granger causality test in the vector error correction framework, he examined the long-term relationship between the two and the direction of their causality [10]. However, these methods face issues of data complexity and algorithm applicability when applied to the construction of causal graphs. Therefore, this article proposes the use of causal reasoning combined with graph theory optimization to address the shortcomings of traditional causal graphs in terms of accuracy and practicality.

3 Methods

3.1 Causality Extraction

3.1.1 Rule Based Approach

Rule based methods rely on pre-defined rule templates to match causal expressions in text. These rules typically include causal conjunctions and specific sentence structures. Although this method is simple and intuitive, its limitation is that the formulation of rules is complex and difficult to cover all situations. In addition, the generalization ability of rule-based methods is poor, making it difficult to handle complex semantic relationships [11].

3.1.2 Statistical Machine Learning Based Approaches

Statistical machine learning based methods utilize classifiers for causality identification by constructing feature engineering. This type of approach can handle more complex text features, but relies on a large amount of labeled data for training. The specific steps are as follows:

Feature extraction: it extracts features such as syntax, vocabulary and context from the text.

Model training: it trains the classifier using labeled datasets.

Relationship recognition: it applies the trained classifiers to recognize causality in new text data.

3.1.3 Methods Based on Deep Learning

Deep learning based methods utilize models such as CNN, Recurrent Neural Network (RNN), and Generic Neural Network (GNN) to automatically learn text features and extract causality. This article uses a pre trained language model BERT to improve the accuracy of causality extraction. The specific steps are as follows:

Data preprocessing: it performs word segmentation and annotation on textual data.

Model training: it utilizes pre trained BERT models for fine-tuning and training causality extraction models.

Relationship recognition: it inputs new text into a trained BERT model, automatically identifying and extracting causality within it.

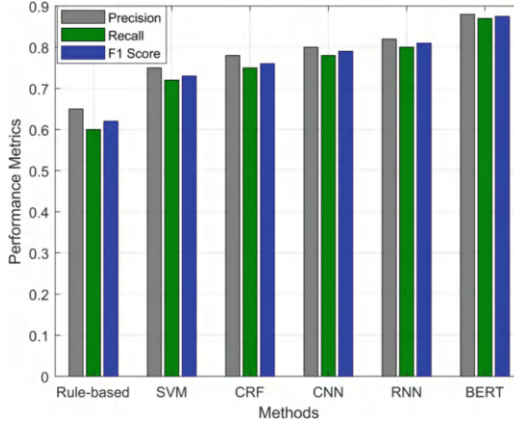


Fig. 1. Comparison of causality extraction methods

In order to evaluate the effectiveness of the BERT model, this study conducted experiments on the SemEval 2010 Task 8 dataset and compared the accuracy, recall, and F1 score of different methods, as shown in Fig. 1:

From Fig. 1, it can be seen that the BERT model based on deep learning outperforms other methods in terms of accuracy, recall, and F1 score, indicating that it performs the best in causality extraction tasks. This further validates the superiority of deep learning methods, especially BERT models, in handling complex text features and semantic relationships.

3.2 Representation of Causality in the Causal Graph

The representation of causality in a causal graph is the core step in constructing an effective graph. This article uses directed edges to represent causality between events, where nodes represent events and directed edges indicate the direction of causality. In this way, the causal chain and logical relationship between events can be visually displayed in the causal graph, making it easier to analyze and understand complex event relationships [12].

Node Description: Each node represents a separate event, and the information of the node includes a description of the event, the time it occurred, and the participants. Providing detailed descriptions of the nodes in the network can help people fully understand the various situations occurring within the network.

Expression of Edges: It is used to express the causality between things. Each edge includes a starting node and an ending node, from “cause” to “effect.” Edges carry weighted or potential information to measure the degree of association and credibility between things. The weights or probabilities can be calculated using the following formula (1):

$$w_{ij} = \frac{C_{ij}}{\sum_k C_{ik}} \quad (1)$$

In formula (1), w_{ij} represents the weight of the edge from node i to node j , and C_{ij} represents the number of occurrences of event j caused by event i .

Weight and probability: In real life, the strength and credibility of influencing factors are crucial information. Weighting can be determined based on the number of occurrences or by expert scoring. Probability was calculated using statistical or machine learning methods. By analyzing the weighted or probabilistic attributes of things, the accuracy of event prediction and decision-making assistance can be improved.

To better demonstrate the representation of causality in a causal graph, we construct a graph using a set of example data. Table 1 lists some events, their causality, and weight information:

Table 1. Partial events, their causality, and weight information

Event ID	Event Description	Occurrence Date	causality	Weight
E1	Machine Failure	2023-01-15	$E1 \rightarrow E2$	0.8
E2	Production Halt	2023-01-16	$E2 \rightarrow E3$	0.9
E3	Order Delay	2023-01-17	$E1 \rightarrow E4$	0.7
E4	Customer Complaint	2023-01-18	$E3 \rightarrow E4$	0.85

Based on the example data in Table 1, we can construct the following causal graph:

Nodes: E1, E2, E3, E4.

Edges: $E1 \rightarrow E2$ (0.8), $E2 \rightarrow E3$ (0.9), $E1 \rightarrow E4$ (0.7), $E3 \rightarrow E4$ (0.85).

This graph clearly displays the causality between various events and their strengths, thus providing a better understanding of their relationships and logical sequence.

On this basis, not only can we clearly display the causality between events, but we can also conduct deeper research using the provided information on weights and probabilities. This method plays a very important role in forecasting, decision support, and knowledge inference.

3.3 Application Scenarios of Causality in Causal Graphs

Causality have a wide range of application scenarios in causal graphs, providing strong support for the analysis, prediction, and decision-making of complex events. Here are several specific application scenarios:

3.3.1 Event Prediction

By constructing a causal graph that includes causality, it is possible to analyze the causal chains and logical relationships between events, thereby predicting possible future events. For example, in supply chain management, understanding the causality between equipment failures (such as E1: machine failures) leading to production shutdowns (such as E2: production shutdowns) and order delays (such as E3: order delays) can predict and respond to potential production disruptions in advance. This helps enterprises optimize resource allocation, improve production efficiency, and reduce operational risks.

3.4 Risk Management

In risk management, causal diagrams can help identify and evaluate potential risk factors and their interrelationships. For example, in the financial industry, by constructing a causal graph and analyzing the impact of economic events (such as E1: economic recession) on market volatility (such as E2: market downturn) and corporate bankruptcy (such as E3: corporate bankruptcy), it can provide a basis for risk prevention and control of financial institutions. By dynamically updating the graph, risk managers can timely grasp the trend of risk changes, formulate corresponding response strategies, and reduce potential losses. The risk assessment is shown in formula (2):

$$R = \sum_{i=1}^n P(E_i) \cdot I(E_i) \quad (2)$$

In formula (2), R represents the overall risk, $P(E_i)$ represents the probability of event E_i occurring, and $I(E_i)$ represents the impact of event E_i .

3.4.1 Complex System Analysis

Causal reasoning diagrams also have important application value in complex system analysis. For example, in intelligent transportation systems, analyzing the causality between traffic accidents (such as E1: traffic accidents) and road congestion (such as E2: road congestion) and travel delays (such as E3: travel delays) can provide scientific decision support for traffic management departments. Based on causal analysis, managers can optimize traffic signal settings, develop emergency plans, improve overall traffic efficiency, and reduce accidents. The traffic flow prediction can be represented by formula (3):

$$Q = K \cdot V \quad (3)$$

Among them, Q represents traffic flow, K represents vehicle density, and V represents average vehicle speed.

In short, causality have a wide range of application scenarios in causal graphs, and their core value lies in the ability to reveal the inherent connections between events, providing scientific basis for the analysis, prediction, and decision-making of complex events. This not only enhances the practicality of the graph, but also provides new ideas and tools for innovation and optimization in various industries [13].

4 Results and Discussion

4.1 Dataset and Experimental Settings

This work used many publicly accessible datasets for experiments, namely the SemEval 2010 Task 8 dataset and the Event StoryLine Corpus dataset, to confirm the application impact of causality in causal networks. A common dataset for determining multiple event links that covers a significant amount of causality annotations is SemEval 2010 Task 8. The Event StoryLine Corpus dataset is a valuable tool for identifying causality and extracting events from real-world stories. These datasets guarantee the scientific validity of the experiments and the diversity of the data by giving us rich textual resources and causality annotations for our study.

A deep learning-based technique for extracting causality was employed in the experiment. To be more precise, we employed pre-trained BERT and CNN and RNN algorithm models to increase the accuracy of causality identification. To further show the causality and logical relationships between events, a causal graph was built during the experiment using the Neo4j graph database. These experimental setups allow us to examine the efficacy and usability of causal reasoning graphs in real-world applications in addition to assessing the performance of causality extraction techniques.

4.2 Experimental Results and Analysis

4.2.1 Performance Evaluation of Causality Extraction

In the causality extraction performance evaluation experiment, we will analyze the causes in the deep network using the SemEval 2010 Task 8 dataset and utilize CNN algorithms, RNN algorithms, and models that incorporate the pretrained language model BERT for causality identification. After the experiment, the metric data for each model were plotted into bar charts to visually display the performance of each model.

In Fig. 2, the F1 value of the BERT model reaches 0.875, which is significantly better than the 0.825 of the CNN model and the 0.825 of the RNN model. In addition, the BERT model achieved accuracy and recall of 0.88 and 0.87, respectively, both higher than the CNN model's 0.85 and 0.80, as well as the RNN model's 0.83 and 0.82. Overall, the BERT model performs the best in causality extraction tasks, significantly improving the model's extraction performance. The specific data is shown in Fig. 2:

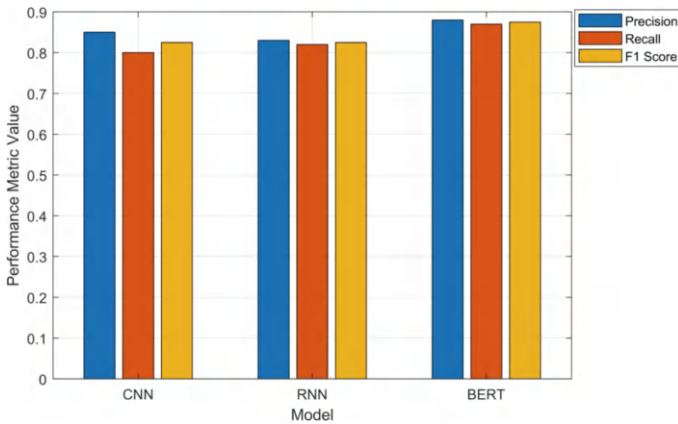


Fig. 2. Performance evaluation of causality extraction

4.2.2 Construction and Optimization of Causal Reasoning Diagram

This experiment uses the Event StoryLine Corpus dataset to extract causality between events through deep learning models, and constructs a causal graph using the Neo4j graph database. Then, the graph is optimized using PageRank and HITS algorithms.

The evaluation indicators include node importance, graph connectivity, and average path length, aiming to evaluate the improvement effect of optimization methods on graph structure and performance. The specific data is shown in Fig. 3.

In Fig. 3, PageRank optimization increased node importance from 0.05 to 0.075, graph connectivity from 0.70 to 0.75, and average path length from 3.5 to 3.2. HITS optimization further increases node importance to 0.08, graph connectivity to 0.85, and shortens average path length to 3.0. These results indicate that the optimized causal graph is more effective and accurate in representing event relationships and logical structures.

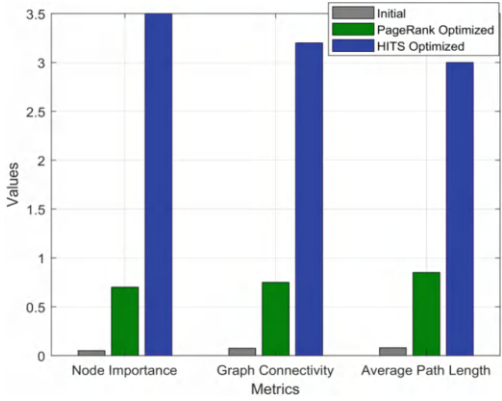


Fig. 3. Construction and optimization of causal reasoning diagram

4.3 Discussion and Improvement

Although the experimental results indicate that causality have important application value in causal graphs, there are still some problems and challenges, such as:

The recognition of implicit causality: Current methods still have shortcomings in the recognition of implicit causality, and further improvement is needed to enhance the semantic understanding and contextual analysis capabilities of the model.

Cross domain applicability: Events and causality in different domains have different characteristics and patterns, and how to construct a cross domain universal causal graph is an urgent problem to be solved.

Dynamic update and expansion: As new events occur and relationships evolve, the causal graph needs to be continuously updated and expanded to maintain its timeliness and accuracy.

In response to the above issues, further exploration can be conducted in the future to develop more efficient implicit causal relationship recognition methods, construct cross domain knowledge transfer frameworks, and implement dynamic update mechanisms for causal graphs.

5 Conclusion

This article studies the application of causality in causal graphs and proposes a method that combines deep learning and graph theory optimization. Firstly, the BERT model is used to effectively extract causality from textual data. Subsequently, an initial causal graph was constructed using the Neo4j graph database, and the graph was optimized using PageRank and HITS algorithms. Experimental results have shown that the optimized causal graph significantly improves node importance, graph connectivity, and average path length, enhancing the graph's ability in event relationship representation and logical structure analysis. However, the method proposed in this article still has shortcomings in identifying implicit causality, and the semantic understanding and contextual analysis abilities of the model need to be further improved. In addition, the applicability of current methods in different fields needs to be verified, and the construction of cross domain universal causal graphs remains a challenge. As new events occur and relationships evolve, the causal map needs to be constantly updated and expanded to maintain its timeliness and accuracy. Future research can explore more efficient methods for identifying implicit causality, construct cross domain knowledge transfer frameworks, and implement dynamic update mechanisms for causal graphs to further enhance their application value and accuracy.

Acknowledgements. This work was supported by 2023 scientific research project of Software Engineering Institute of Guangzhou, Project No: KY202308.

References

1. Zhu, Y., Yi, J., Xie, J., et al.: Deep causal reasoning for recommendations. *ACM Trans. Intel. Sys. Technol.* **15**(4), 1–25 (2024)
2. Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(10), 11624–11641 (2023)
3. Liu, Y., Wei, Y.S., Yan, H., et al.: Causal reasoning meets visual representation learning: a prospective study. *Mach. Intel. Res.* **19**(6), 485–511 (2022)
4. Shiying, Z., Yang, L.: An empirical study on the support path of emergency event intelligence by integrating the knowledge graph of affairs and network public opinion analysis: taking hazardous chemical accidents as an example. *J. Info. Res. Manage.* **13**(4), 60–71 (2023)
5. Biao, A., Xingfen, W.: A method for constructing a trend graph of bulk agricultural product prices. *J. Beijing Univ. Info. Sci. Technol. Natu. Sci. Edit.* **38**(5), 25–31 (2023)
6. Deng, Z., Weng, D., Xie, X., et al.: Compass: towards better causal analysis of urban time series. *IEEE Trans. Visual Comput. Graphics* **28**(1), 1051–1061 (2021)
7. Yue, C., Yuhao, H., Yawei, S., et al.: A causal short answer solving method based on abstract reasoning graph. *Chinese J. Info. Sci.* **36**(4), 124–136 (2022)
8. Yang, J., Han, S.C., Poon, J.: A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.* **64**(5), 1161–1186 (2022)
9. Yang, Z., Wang, M.C., Chang, T., et al.: Which factors determine CO2 emissions in China? trade openness, financial development, coal consumption, economic growth or urbanization: quantile granger causality test. *Energies* **15**(7), 2450–2461 (2022)

10. Wesiah, S., Onyekwere, S.C.: The relationship between financial development and economic growth in the United Kingdom: a granger causality approach. *Quantitative Econ. Manage. Stud.* **2**(1), 47–71 (2021)
11. Liang, X.S.: Normalized multivariate time series causality analysis and causal graph reconstruction. *Entropy* **23**(6), 679–684 (2021)
12. Whisman, M.A., Sbarra, D.A., Beach, S.R.H.: Intimate relationships and depression: Searching for causation in the sea of association. *Annu. Rev. Clin. Psychol.* **17**(1), 233–258 (2021)
13. Mathlin, G., Freestone, M., Jones, H.: Factors associated with successful reintegration for male offenders: a systematic narrative review with implicit causal model. *J. Exp. Criminol.* **20**(2), 541–580 (2024)



Research on Machining and Simulation Optimization System of Automobile Steering Knuckle Based on Advanced Algorithm

Xiaopeng Chang¹, Siyu Chen¹, Xiyu Zhang¹, Bangcheng Zhang², and Bo Yu²(✉)

¹ School of Mechanical and Electrical Engineering, Changchun University of Technology, Changchun 130012, China

² School of Mechanical and Electrical Engineering, Changchun Institute of Technology, Changchun 130012, China
yubo745@163.com

Abstract. With the rapid development of automobile industry, automobile steering knuckle is one of the most important parts on automobile steering bridge, therefore, the design of its high-efficiency production process has become the inevitable choice for the development of automobile steering knuckle. In view of the characteristics of automobile steering knuckle, such as complex structure, many machining parts, high machining requirements, difficult positioning and clamping, and large production batch, combined with the actual machining process of automobile steering knuckle, including production line layout and position arrangement, production rhythm and production capacity status quo, discussed and analyzed the production line of our car steering knuckle specific process problems. According to the structure characteristics of automobile steering knuckle, the machining technology scheme is worked out, and the machining positioning datum, part clamping scheme and balance analysis are given. Based on the Quest platform, the simulation test of maximizing the machining efficiency and optimizing the man-hour configuration is carried out, and the simulation results are analyzed and evaluated, the validity of the research is proved. Finally, the optimal process plan of the automobile steering knuckle is given, which is helpful to the reference design of the machining process.

Keywords: Automobile Industry · Steering Knuckle · Processing Technology · Optimization Design · Simulation Analysis

1 Introduction

With the rapid development of our country's manufacturing industry and the continuous increase of the holding of Volkswagen, the automobile industry has entered a new chapter. The steering knuckle of an automobile under running conditions bears a variable impact load, and supports and drives the wheels to rotate around its main pin to make the automobile turn [1, 2], therefore, the automobile steering knuckle has become one of the important parts on the automobile steering bridge.

In actual production, the automobile steering knuckle has the characteristics of complex structure, many machining parts, high machining requirements, difficult positioning and clamping, and large production batch, etc., at the same time, the automobile steering knuckle has a large number of dimensional tolerance requirements, geometric tolerance requirements and spatial multi-angle correlation hole system, because the processing technology can only adapt to its structural characteristics, the method of scattered processing procedure increases the cumulative error of repeated clamping, increases the probability of accidental error, and reduces the guarantee ability of product consistency, therefore, the design of automobile steering knuckle high-efficiency machining process optimization, fixture development and application has become the current research hot issues [3].

Taking automobile steering knuckle as the research object and combining with the actual production capacity demand of automobile steering knuckle, aiming at the optimization of the high-efficient processing technology of automobile steering knuckle and the development and application of the fixture, complete the optimization of high-efficient processing technology, fixture development and application, to improve product quality and processing efficiency, to meet the high-efficiency, high-precision, large-scale processing needs. This study has a strong engineering practice significance for solving the practical needs of automobile steering knuckle processing, and has a guiding significance for the high-efficiency and high-precision machining of special-shaped space structure.

2 Related Works

Due to the complex spatial structure and more spatial machining features of automobile steering knuckle, the optimization analysis and simulation design of its machining process have become a hot issue in the current manufacturing field, many researchers have carried out in-depth research on this issue, but also made some research results.

Literature [4–6] has carried out continuous research on the dimensional tolerance requirements of automobile steering knuckle, the features of spatial multi-angle hole series and the feasibility points of the process decentralized machining method, and has obtained certain results, its research content has the strong practicability, also may lay the foundation for the automobile steering knuckle processing technology thorough analysis and the simulation research.

Literature [7, 8] has carried on the thorough analysis according to the automobile steering knuckle production capacity demand, the production characteristic, the product characteristic and the production line layout, has carried on the effective analysis through to the production data, the basic data of automobile steering knuckle manufacturing are obtained, which points out the direction of automobile steering knuckle high-efficiency production.

Literature [9, 10], in view of the present situation of automobile steering knuckle processing, combined with automobile steering knuckle positioning, clamping, balance analysis, etc., the process design of automobile steering knuckle is simulated and realized by integrating Quest optimization analysis software and Solidworks 3D design software, which provides a new method for the research of automobile steering knuckle process.

3 Methods

3.1 Analysis of Processing Characteristics

There are a lot of dimensional and geometric tolerances in automobile steering knuckle, and they form a complicated dimensional chain because of the multi-orientation of the steering knuckle and the multi-angle of the holes [3]. The specific size requirements of the steering knuckle to be processed are shown in Figs. 1, 2 and 3.

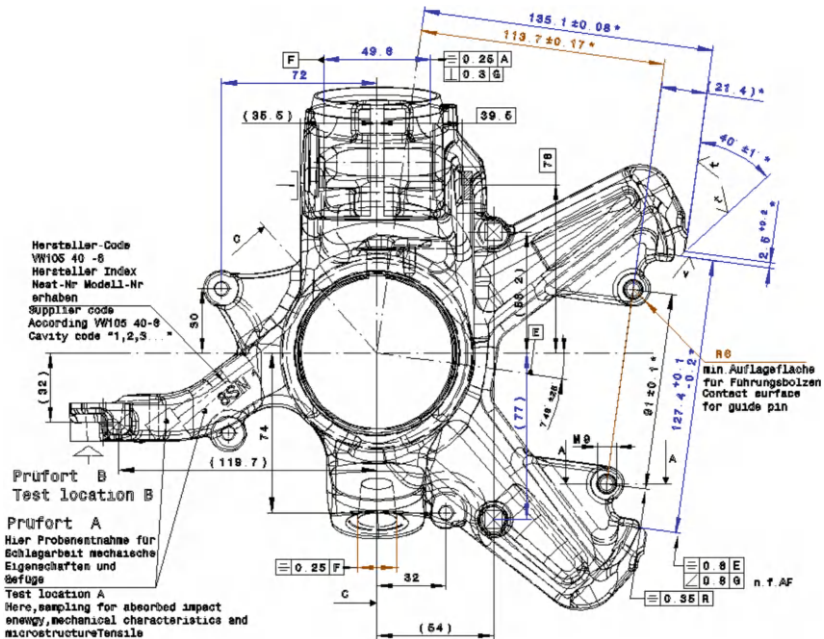


Fig. 1. View 1 of machining automobile steering knuckle

Because of the high requirement of many machining technologies, the special fixture must be considered reasonably in the choice of coarse datum, orientation mode and clamping mode to ensure its high efficiency, high-quality batch processing of the parts.

The production process of the automobile steering knuckle mainly for machining part, so the process involved here is machining process. Among them:

- (1) the inspection shown in the processing process diagram is periodic inspection, the hole after processing needs 5 inspection, inspection once, surface processing needs 20 inspection once, mainly for the production self-inspection process;
- (2) the quality department regularly inspects the quality of the products;

- (3) after the adjustment of the cutting tools or the replacement of the jigs and fixtures, the previous products must be processed and sent to the quality department for inspection.

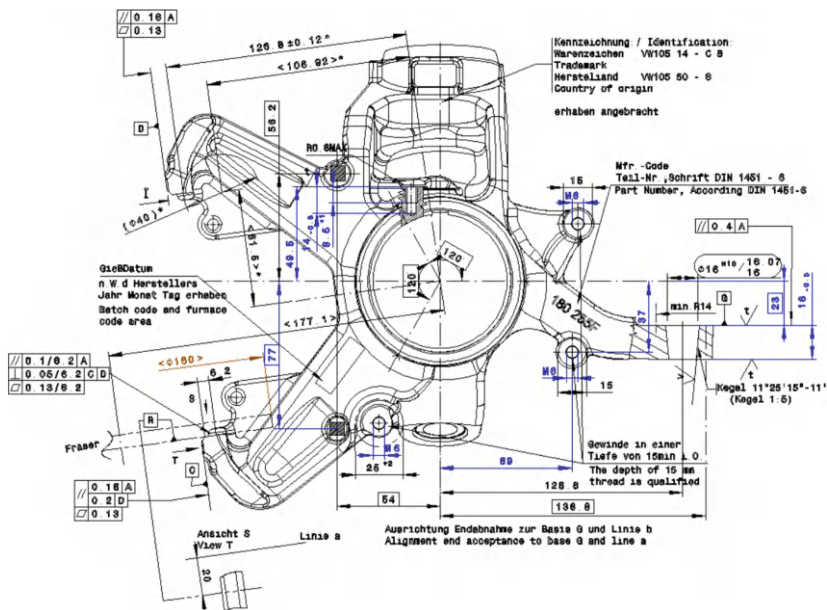


Fig. 2. View 2 of machining automobile steering knuckle

3.2 5W1H and ECRS Analysis

Table 1 shows the 5W1H and ECRS analysis tables for automobile steering knuckles.

3.3 Production Line Layout and Working Station Arrangement

The layout of the steering knuckle production line is shown in Fig. 4. From the layout of the machining department, it can be seen that the whole production line is approximately a U-shaped distribution. In the logistics process, the work-in-process is carried out by 40 first-rate carts, operating staff in processing a flow after the car will be transported to the next station, the distribution of the machine tools processed by the operator is also shown in the diagram, a total of 10 stations.

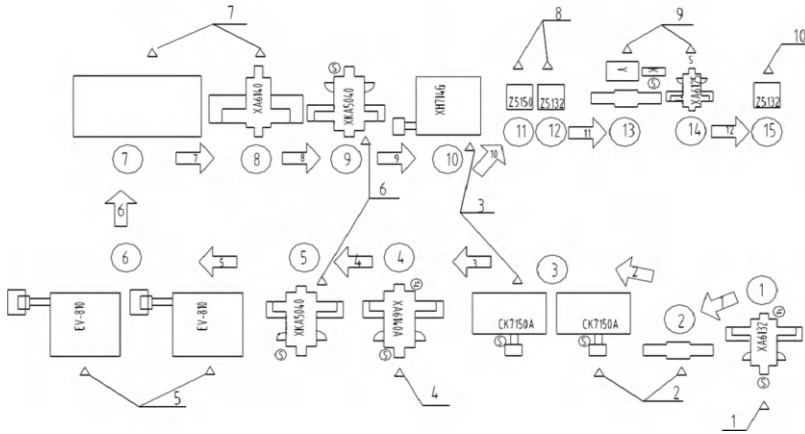


Fig. 4. Production line layout diagram of automobile steering knuckle

3.4 Capacity Analysis

3.4.1 Production Tempo

The above analysis: Milling Face 26 s/piece, tapping thread 25 s/piece, drilling center hole 49.75 s/piece, rough and fine turning 99.5 S/piece, milling positioning face 51 s/piece, milling brake face 118 s/piece, milling plane, drilling, chamfering angle 112.75 s/piece, rough milling two ears four sides 68 s/piece, semi-finish milling two ears four sides 68 s/piece, finish milling two ears four sides 68 s/piece, milling joint arm two sides 118 s/piece, drilling tapered hole of hinge arm 99.5 s/piece, drilling tapered hole of long end of Hinge 79 s/piece, rough drilling tapered hole of short end of Hinge 79 s/piece, fine boring tapered hole 42 s/piece, milling keyway 42 s/piece.

From the above data, the milling brake surface is 118 s/piece, in which the milling brake surface and the milling joint arm on both sides of the work station, the work-taking unit recorded in the work measurement is the time when the operator stands beside the machine and begins to take the work-piece, but in the actual production process, before the workpiece on both sides of the milling joint arm is clamped by the operator, the machine tool for milling the brake surface has stopped, and the average time is 3.0 s, the average time for the operator to move from both sides of the milling arm to the milling brake surface is 2.5 s.

3.4.2 Capacity Analysis

The production time is 126 S/piece. The normal production time of the machining department is 8:00 am-16:45 PM. The lunch break is from 11:15 to 12:00. In addition to the normal production time in the morning, about half an hour after the change of tools, adjustment of machine tools and other trial production process, this period of time in the 2.4.1 job determination method and calculation formula did not take into account the scope of relaxation. Therefore, considering the above, the actual production time of a production line shift is 7.5 h, without considering the relaxation and excluding the human and environmental impact, the maximum output of a shift calculated using normal time

was $7.5 \times 3600/126 = 214.3/\text{shift} \approx 214/\text{shift}$, if 21% relaxation was considered, it was $7.5 \times 3600/(126 \times (1 + 21\%)) = 177.1/\text{shift} \approx 177/\text{shift}$. As a result, the maximum day shift capacity can reach 214, the minimum capacity of 177.

3.5 Summary of Production Line Problems

Through observing the production line, combining with the analysis of process procedure and 5W1H question and ECRS, some problems in the actual production process are summarized as follows:

- (1) the amount of work-in-process is too large and the capital is too large, and it needs to adopt three shifts for production, and three shifts will greatly increase the fatigue strength of workers;
- (2) the unreasonable synchronization of working procedures and the long walking distance will lead to the excessive fatigue strength of operators, which will affect the processing efficiency, but part of the work station is more idle, the operator idle time is more, part of the work station is basically no idle, time, collocation is not reasonable;
- (3) in the actual production process, because collocation is not reasonable, the process synchronization is not reasonable, by the operators themselves, adjust the order of operation, not in accordance with the process of strict processing work, resulting in on-site processing sequence confusion.

4 Results and Discussion

4.1 Positioning Reference

The steering knuckle part is a special-shaped part with many space arms, and the production batch is large [11]. The processing of coarse datum is not suitable for the conventional processing mode of line-drawing and alignment, and because many holes and faces are located on the side arm far away from the center hole, if there is a deviation of the rough datum position, it will lead to the scrap of the parts directly. Therefore, the plane positioning of three points in space is considered. Because the three points in space are in the same casting box, the inevitable closing error can be reduced. The adjustable clamping device of two mutually perpendicular bending arms is used to realize the center adjustable positioning, and the two side baffles of the opposite orientation platen are also used for auxiliary alignment.

4.2 Clamping Scheme

The clamping surface of the workpiece is an irregular plane, and the holes and faces to be machined not only have very high form and position tolerance requirements. But also has the complex spatial dimension chain requirements. In order to ensure the precision of the workpiece, reduce the cost and improve the production efficiency, the side arm of the special-shaped plate is used.

4.3 Equilibrium Analysis

Due to many factors such as uneven material, defects of casting blank, machining and assembly errors, asymmetrical geometric shape of design and so on, noise and vibration are produced in high-speed rotating machining of fixture, accelerated tool wear, shorten the life of the machine tool and fixture, serious may cause destructive accidents. Therefore, it is necessary to carry out counterweight balance calculation on the fixture to make it reach the allowable level of balance precision or to reduce the mechanical vibration amplitude to the allowable range.

First of all, solidworks 3D design software, which has rich parts modeling function, is used to model the parts of fixture. Second, it enters the assembly mode to complete the virtual assembly, the interference detection is used to ensure the design is reasonable and reliable.

4.4 Quest Introduction

Quest is a full 3-d digital factory environment for DELMIA to simulate and analyze the accuracy and efficiency of production processes. Quest-based simulation environment combines powerful visualization and import/export functions, making it the preferred solution for engineering and management of production process simulation and analysis. In Quest, simulation statistics can be displayed directly or saved as files for other software to use directly. In QUEST, the system provides a number of standard report templates for simulation results, as well as very convenient report customization capabilities. Statistical data can be displayed in tables, graphs in various forms, such as pie charts, histograms, linear charts, etc.. By modifying the parameters and changing the model, we can observe and compare the different behaviors and results of each simulation. Quest automatically captures the results of each simulation run and can compare and cross-analyze the results of multiple simulations that users require.

4.5 The Idea of Simulation

If we want to understand the personnel arrangement, the arrangement of equipment and the working condition of the operators after the design of the new project, we can make use of the Quest simulation software, modeling and simulation of production line system. Quest can clearly represent the production status of the machinery plant, so the use of this software for simulation. In the simulation, we can clearly see the production situation and the position of the bottleneck short board, so we can improve the bottleneck position. After the improvement, we can simulate the production line system again,

observe whether there will be a bottleneck again, and then for the bottleneck of the station to improve, to achieve continuous optimization of the production line improvement. Because there are three new schemes studied, the design ideas and concepts involved are different, the side and emphasis of each scheme are also different, how to let the decision-maker choose the best scheme is also what Quest needs to complete, after the Quest system runs the design, the decision maker can see the production line at a glance from the software. At the same time, due to the design of the content of different emphasis can also be based on the current production requirements to choose.

4.6 Simulation Setup

Because the simulation can not completely restore the production line, the actual production line is more complex than the simulation, because in the actual production process, part of the processing, testing and other content can not be realized in the simulation, so in the simulation, the use of data to represent these can not be expressed out of the content, but most of the data consistent with the current production situation. The following are several settings:

- (1) the time for clamping and unloading the workpiece according to the time requirements in the work measurement;
- (2) the time for the equipment to process the workpiece according to the time requirements in the work measurement;
- (3) taking and placing the workpiece on the conveyor belt according to the time requirement in the work measurement;
- (4) in the actual work, the workbench is replaced by Buffer in the simulation;
- (5) the time for the workpiece to be deburred and inspected on the workbench is replaced by the time Labor stays on Buffer in the simulation;
- (6) the number of workpiece transfers on the Conveyor is four, as part of the Conveyor is not possible due to the shorter route, there are 4 workpieces placed, and the actual position is 3, but it does not affect the processing speed and other workpieces, the actual arriving workpieces are still 4.
- (7) the grey workpieces in the simulation are Jac shaft steering knuckle;
- (8) the workpiece in the simulation can reach the next work station on the guide rail without human action, in reality, the process must push the iron plate trolley to the next work station by human action.

4.7 Simulation Implementation

The simulation is based on an improved production line that maximizes efficiency and optimizes man-hours, as shown in Fig. 5, for the specific operation and detailed layout can refer to the Quest simulation program attached to this paper, in which, there will be a specific program implementation process for the practical observation of decision-makers.

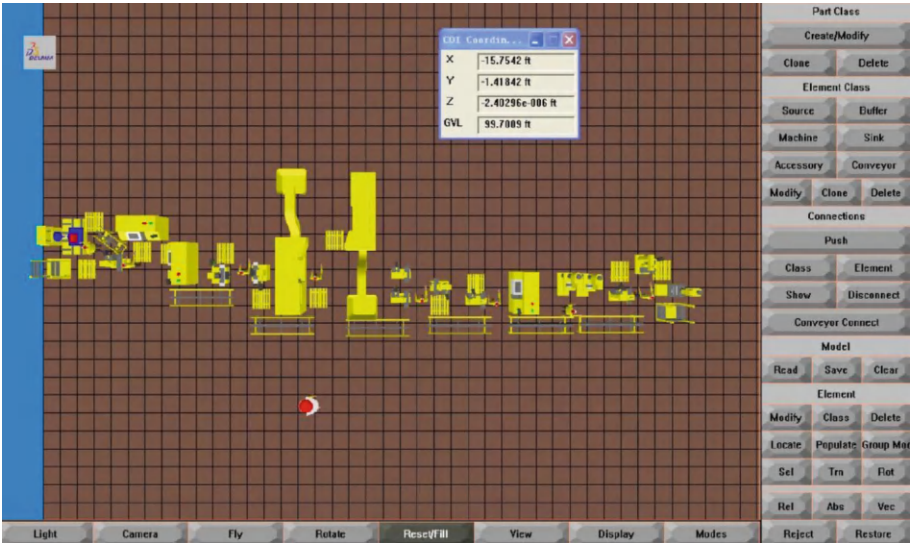


Fig. 5. Based on an operational diagram that maximizes efficiency and optimizes man-hours

4.8 Simulation Analysis

The goal of this simulation analysis is the utilization of personnel, different programs for the utilization of personnel in different places, Quest simulation software, it can show the man-machine utilization rate of the simulation to the readers, therefore, it can analyze the man-machine utilization rate by this way, of course, in the analysis process must be combined with accurate data to achieve a real-time observation purposes. The simulation results are shown in Fig. 6. Through the real-time observation of man-machine utilization in Quest simulation, it can be clearly seen that in the improved design scheme, the utilization rate of personnel in milling face, drilling center hole and rough and fine turning station is the lowest, but it still reaches 47% utilization rate, and the utilization rate in other stations is more than 58%, which is worth emphasizing, the improved production line is due to the improvement of the model of 10 stations and 8 workers who can not produce at the same time to the model of 7 stations and 7 workers who can produce at the same time, from the root to eliminate such as the long and short end of the drill reaming cone hole station, and tapping thread station 100% of the personnel utilization. The above staff utilization situation also appears, is precisely because the production line based on the efficiency maximization improvement result, cuts a person to cause each station staff utilization to increase.

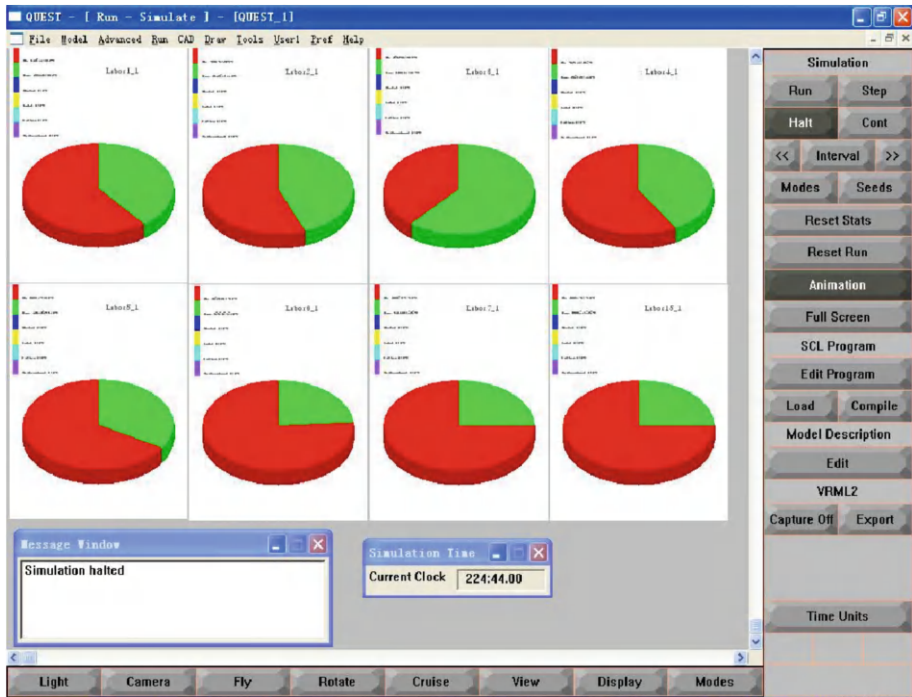


Fig. 6. The real-time observation chart of the integrated design scheme

4.9 Simulation Evaluation

Quest simulation is a good way to show the production system studied, because the machine tool style is similar or identical to the machine tool in the actual machine shop, and can be presented to the reader clearly, at the same time can reflect the details of the problem, for the machine tool parameters set, you can set the processing time, clamping time, unloading time, but also in the actual production process, can explain the operator's actual operation of the machine tool sequence, as well as the path of personnel walking, so as to guide the operator in accordance with the simulation of the processing sequence of the work station processing, reducing unnecessary walking, the utilization rate of personnel has been further improved. Of course, there are some drawbacks in simulation, because the simulation software can not completely reflect all aspects of the production line, after all, the actual production site there are always many uncertainties, including some settings, it was also the result of Quest not being able to fully reflect the production line, so some other settings were used instead.

5 Conclusion

After the above simulation and practice analysis, the most form of the automobile steering knuckle processing process for the completion of two procedures, the specific processing as shown in Table 2, the required fixture as shown in Figs. 7 and 8.

Table 2. Optimal machining procedure of automobile steering knuckle

Num	Processing content	Processing technology requirements
1 preface	Boring bearing holes	$\phi 73.919\text{--}\phi 73.949$; $\phi 70\text{--}\phi 71$; $\phi 77\text{--}\phi 77.3$; $\phi 74.1\text{--}\phi 74.4$; $\phi 71.85\text{--}\phi 72.2$; $\phi 75.85\text{--}\phi 76.2$; 44 ± 0.1 ; 53.5 ± 0.2 ; 48 ± 0.2 ; 2.15 ± 0.02 ; 2.29 ± 0.02 ; 42.3 ± 0.05 ; Rz16
	Milling 2-M9 plane, drilling, tapping 2-M9	$\phi 9 \pm 0.3$; 0.18 ± 0.05 ; 2-M9; 53.4 ± 0.1 ; 91 ± 0.1 ; 13.7 ± 0.17 ; $7^\circ 48' \pm 25'$
	Machining 3-M6 holes and planes	$\phi 7 \pm 0.30$; 1.6 ± 0.05 ; 3-M6; 25.3 ± 0.2 ; 32.3 ± 0.2 ; 18.5 ± 0.2
	Hinge 16 taper hole and milling plane	$\phi 16 \pm 0.07$; $11^\circ 25' 15'' \pm 11'$; 126.8 ± 0.2 ; 81.5 ± 0.2 ; 23 ± 0.1 ; $18\text{--}0.5$
	Hinge 18 taper hole and milling plane	$\phi 18 \pm 0.07$; $11^\circ 25' 15'' \pm 11'$; 54.1 ± 0.2 ; $21^\circ \pm 15'$; 77.3 ± 0.2 ; $\phi 26 \pm 0.2$; 1.5 ± 0.2 ; $18.01\text{--}0.5$
	Bore the shock absorber hole	$\phi 49.6 \pm 0.025$; 130.2 ± 0.2 ; 120.5 ± 0.2 ; $8^\circ 3' \pm 7'$; Rz3 ± 0.2 ; Rz25-Rz40
2 preface	Milling plane	135.1 ± 0.08 ; $40^\circ \pm 1^\circ$; $7^\circ 48' \pm 25'$
	Slotting and drilling	$6.350\text{--}0.55$; 23 ± 0.3 ; 0.67 ; 24 ± 0.3 ; 0.67
	Milling arc surface	127.1 ± 0.1 ; 0.2 ; 127.1 ± 0.1 ; 0.2 ; 6.2 ± 0.2 ; 2.5 ± 0.2 ; $7^\circ 48' \pm 25'$
	Milling ABS surface, drilling and reaming $\phi 10$ hole and tapping M 6	M6; 8.5 ± 1 ; $14\text{--}0.5$; $45^\circ \pm 1^\circ$; $\phi 10 \pm 0.036$; $\phi 22 \pm 0.2$

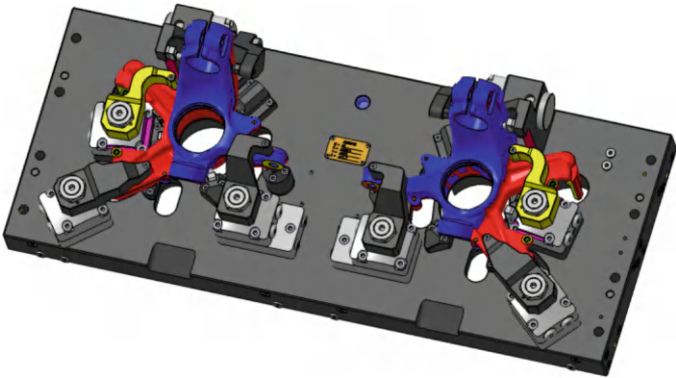


Fig. 7. First sequence jig drawing

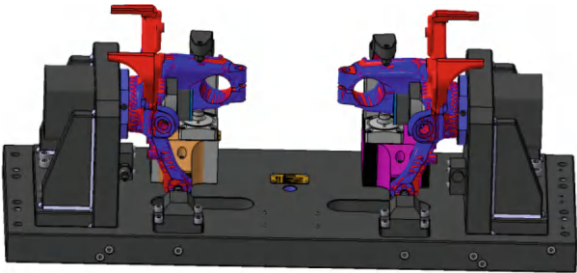


Fig. 8. The second processing sequence jig drawing

- (1) The function of automobile steering knuckle in passenger cars is analyzed, and the necessity of optimization design of automobile steering knuckle processing technology is pointed out;
- (2) Aiming at the characteristics of the current automobile steering knuckle, such as complex structure, many machining parts, high machining requirements, difficult positioning and clamping, large production batch, etc., the difficulties in the process optimization design are pointed out;
- (3) The process route of the steering knuckle is analyzed and the relevant optimal process route is worked out;
- (4) Based on the Quest platform, the comprehensive simulation analysis and evaluation of the machining process were carried out, and the validity of the research content was demonstrated, the implementation plan of the optimal machining process is given, which can be used as a reference for the machining process design of related special-shaped structure parts.

Acknowledgments. This work is supported by the General Project of Free Exploration of Science and Technology Department of Jilin Province, No. 20240302050GX.

References

1. Bratcu, A.I., Dolgui, A.: Some new results on the analysis and simulation of bucket brigades (self-balancing production lines). *Int. J. Prod. Res.* **47**(2), 369–387 (2022)
2. Shah, R., Peter, T.W.: Defining and developing measures of lean production. *J. Operat. Manage.* **25**(4), 785–805 (2017)
3. Huang, H.-C., Lee, L.-H., Song, H., et al.: SimMan—A simulation model for workforce capacity planning. *Comp. Oper. Res.* **36**(8), 2490–2497 (2022)
4. Nakade, K., Nishiwaki, R.: Optimal allocation of heterogeneous workers in a U-shaped production line. *Comp. Indus. Eng.* **54**(3), 432–440 (2023)
5. Miltenburg, J.: One-piece flow manufacturing on U-shaped production lines: a tutorial. *IEE Transactions* **33**(4), 303–321 (2021)
6. Kara, Y., Paksoy, T., Chang, C.-T.: Binary fuzzy goal programming approach to single model straight and U-shaped assembly line balancing. *John Wiley & Sons. Inc.* **195**(2), 335–347 (2018)
7. Tu, C.-S., Chang, C.-T.: Using binary fuzzy goal programming and linear programming to resolve airport logistics center expansion plan problems. *Applied Soft Computing* **44**, 222–237 (2023)
8. Lee, H.-T., Michael, H.W.: On the search of workstations arrangement in pull production systems. *Comp. Indus. Eng.* **54**(3), 613–623 (2023)
9. McMullen, P.R., Frazier, G.V.: Using simulated annealing to solve a multiobjective assembly line balancing problem with parallel workstations. *Int. J. Prod. Res.* **36**(10), 2717–2741 (2020)
10. Saurin, T.A., Ferreira, C.F.: The impacts of lean production on working conditions: a case study of a harvester assembly line in Brazil. *Int. J. Indus. Ergon.* **39**(2), 403–412 (2023)
11. Denise, R., Dirk, B., Van, L.H. et al.: Impact of lean production on perceived job autonomy and job satisfaction: an experimental study. *Human Fact. Ergonom. Manuf. Ser. Indus.* **26**(2), 159–176 (2016)



Network Partitioning and Demand Characterization for Management of Urban Low Voltage Power Distribution Systems

Haisheng Hong^(✉), Yongshu Chen, Zhifang Zhu, Zheng Sun, Jiarui Guo, and Qin Lin

Guangzhou Power Supply Bureau of Guangdong Power Grid Co., Ltd., Guangzhou, China
honeyhycere@qq.com

Abstract. The power distribution network is considered complex and hence the reliable and safe operation is a non-trivial task. At present, grid management of urban power grids is an important and complex task in modern urban management. This provides the opportunities for more efficient management of the power distribution systems. This work exploited the network partitioning and demand characterization method for management of low voltage power distribution systems. In detail, this work investigated the evaluation and application strategies of power grid in urban villages, and constructed a grid indicator system from five dimensions: power supply capacity, equipment operation, power quality, equipment level, and intelligence level. In addition, the electricity demand is characterized through data analysis. Finally, the proposed solution is evaluated through a case study and the numerical results confirmed its effectiveness.

Keywords: Risk Assessment · Electricity Power Distribution Networks · Fault Analysis · Operational Scenario Forecasting

1 Introduction

With the acceleration of the global energy transition and the promotion of China's "dual carbon" goals, the integration of distributed energy such as photovoltaics and wind power into urban distribution networks is constantly increasing. This not only enriches the energy structure of the power grid but also puts forward higher requirements for the acceptance capacity and consumption efficiency of the distribution network. The infrastructure construction of urban distribution networks continues to strengthen, including the construction and renovation of power grid lines, substations, distribution rooms, etc. Especially in new urban areas and expansion zones, the construction of the power grid is constantly increasing to meet the growing electricity demand. The transformation and upgrading of the distribution network is also steadily advancing, improving the operational efficiency and reliability of the power grid through measures such as replacing old equipment and optimizing the power grid structure.

The problems in the management of power distribution networks in large cities are complex and multidimensional issues, which not only affect the stability of urban power

supply but also directly impact the quality of life of residents and the normal operation of the social economy. With the acceleration of urbanization, the population density and electricity consumption in big cities have increased sharply. However, the planning of distribution networks often fails to fully anticipate this change, resulting in insufficient capacity of some regional power grids to meet the electricity demand during peak hours. Due to insufficient coordination between urban planning and power grid planning, as well as limitations in funding, technology, and other aspects, the construction of distribution networks often lags behind the development speed of cities, especially in new urban areas and expansion areas, where power grid construction is difficult to keep up with the pace of urban construction.

Equipment in the distribution networks of some large cities has been in operation for many years and has serious aging problems, such as small wire cross-sections, high resistance, and insufficient transformer capacity. These problems not only increase power losses but also reduce the quality of the power supply. Due to limitations in funding, manpower, and other resources, maintenance investment in the distribution network is often insufficient, resulting in frequent equipment failures and affecting the stable operation of the power grid. At the same time, there is a lack of sufficient financial and technical support for the renovation and upgrading of old equipment.

There is often an uneven distribution of electricity load in the current urban distribution network, with some areas having excessively high electricity loads while others have relatively low loads. This uneven load distribution not only increases the operating pressure of the power grid, and can easily lead to local power supply shortages. Meanwhile, in recent years, with the continuous growth of electricity demand, the supply-demand contradiction in the distribution network has become increasingly prominent. Especially during peak electricity consumption periods such as high temperatures in summer and heating in winter, the power supply pressure on the grid further increases, making it easy to experience power shortages.

At present, grid management of urban power grids is an important and complex task in modern urban management. It divides the city into several grid areas to achieve refined, efficient, and intelligent power grid management. The traditional power grid management model is inadequate in dealing with large-scale and complex power grid systems, making it difficult to ensure the safety, stability, and efficient operation of the power grid. The introduction of grid management mode is precisely to solve these problems and improve the efficiency and quality of power grid management. Grid-based management divides the city power distribution systems into several grid areas, each with a dedicated management team and responsible person, achieving refined management and rapid response to the power grid. This management model helps to timely detect and solve problems in the operation of the power grid, improve the reliability and stability of the power grid, and also help optimize resource allocation and reduce operating costs. As a result, the appropriate partitioning and grid-based management of urban power grids is considered an effective means to improve the management efficiency and quality of urban power grids. Significant achievements have been made in shortening the time for handling power grid faults, improving power supply reliability, and increasing customer satisfaction. Grid management also helps optimize the allocation of power grid resources, reduce operating costs, and promote the sustainable development of the power industry.

On the other hand, the characterization of power demand is considered important to guarantee the reliable and safe operation of urban power distribution systems for a number of reasons. The demand characterization can be directly related to the safety, stability, economy, and sustainable development of the power grid. The analysis of power load characteristics can reveal the load variation patterns of the power grid at different times and spatial scales, which helps to timely discover and solve potential problems in the operation of the power grid, such as overload, low voltage, etc., thus ensuring the safe and stable operation of the power grid. Through in-depth analysis of the characteristics of power load, it is possible to predict the future load demand of the power grid, provide the basis for power grid planning and construction, ensure the reliability and stability of power supply, and meet the electricity needs of urban development and residents' lives. The analysis of power load characteristics helps to understand the differences in power demand in different regions and periods, providing data support for the optimal allocation of power resources. By rational allocation of electricity resources, energy utilization efficiency can be improved and electricity production costs can be reduced. With the development of smart grids, the analysis of power load characteristics plays an important role in the intelligent construction of power grids. By real-time monitoring and analysis of load characteristics, intelligent scheduling and management of the power grid can be achieved, improving the automation and intelligence level of the power grid.

To this end, this work develops an efficient solution for network partitioning and demand characterization for the management of urban low-voltage power distribution Systems. The main technical contributions made in this work are summarized as follows:

- (1) This study investigated the evaluation and application strategies of power grid in urban villages, and constructed a grid indicator system from five dimensions: power supply capacity, equipment operation, power quality, equipment level, and intelligence level.
- (2) This work investigated the automatic partitioning method of power professional grid based on government grid. In addition, a genetic algorithm-based method for splicing government basic grids was proposed.
- (3) This study proposes an analysis method for substation load data and user electricity data, and constructs a substation and user load characteristic analysis model.

The rest of this paper is organized as follows: Sect. 2 formulates the problem and presents the proposed solution. The experiments are carried out and numerical results are provided in Sect. 3. Finally, the conclusive remarks are given in Sect. 4.

2 Related Work

The research outlined in [1] underscores the pivotal role of uncertainties in shaping the approach to selecting and designing distributed renewable energy sources alongside battery storage systems. It commences with a comprehensive uncertainty characterization, detailing the input variables influencing the energy planning model. Subsequently, an uncertainty analysis delves into how these variables impact the model's output variability. The authors in [2] emphasize the significance of identifying customers based on seasonal load fluctuations, which not only influence network operations but also enhance resource

planning strategies. In response, we introduce a novel algorithm designed to detect and categorize electricity consumers exhibiting weekly seasonal load patterns, pinpointing the specific weekdays where such patterns are evident. Meanwhile, [3] presents preliminary findings from a comprehensive survey targeting residential load users in an urban setting. This endeavor forms part of a broader collaboration between the University of Padova and AGSM, focused on analyzing the Medium and Low Voltage electrical distribution networks. The collaboration aims to characterize current loads, while envisioning realistic future scenarios that align with sustainable energy policies, such as promoting increased electricity usage in final energy consumption and advancing electrical mobility. In [4], a multifaceted approach utilizing K-means clustering, normalization techniques, and statistical analyses is employed to uncover the distinct features of electricity users. This exploration encompasses industry characteristics, customer value, and electricity demand, mining hidden insights and potential value from consumption data. The outcome is the identification of four archetypal user electricity loads characteristic curves. Lastly, [5] delves into an innovative methodology for characterizing domestic electricity demand in smart cities, leveraging statistical data captured at half-hourly intervals. This approach offers an alternative lens to understand and predict electricity consumption patterns, underscoring the importance of data-driven insights for smart city planning.

The work in [6] first proposes a novel and efficient hierarchical spectral clustering-based network partitioning algorithm followed by a decentralized compressive sensing (DCS)-based state estimation. In [7], a dynamic partitioning method for real-time optimal scheduling of the distribution network is proposed, considering the dynamic reactive power regulation characteristics of distributed power sources such as photovoltaics and energy storage. The work in [8] proposes a parallel restoration model for distribution networks via a shortest-path-based partitioning approach. A distribution network partitioning method based on the shortest path algorithm is introduced, which filters out the unimportant load nodes on the edge by means of power verification. A partitioning method of ADNs by using spectral clustering algorithm is proposed in [9] to avoid the curse of dimensionality and accurately obtain the partitioning scheme of the ADNs. In [10], a dynamic islanding partition method that considers the volatility of DG output and load demand is proposed for the distribution network containing DG and energy storage devices.

3 Problem Formulation and Solution

3.1 Principles and Indicators of Network Partitioning

At present, the grid partitioning of urban distribution networks is mainly based on medium voltage distribution networks. The distribution network planning carried out in hierarchical zoning has the characteristics of clarity, differentiation, and orderliness. However, due to the weak historical low-voltage network and limited land resources, urban villages have uneven and insufficient development characteristics and have not yet formed a scientific and effective hierarchical zoning method. The planning and construction of the urban village distribution network focus on solving the power supply quality problems of individual substations, which have problems, e.g., lack of coordination with the development of the urban power grid, low accuracy of load forecasting, weak power

supply reliability, and insufficient investment precision. With the continuous promotion of low-voltage transparency in the distribution network, the government adopts grid management to carry out social governance. The distribution network of urban villages also urgently needs to be managed through digital and grid-based methods for hierarchical zoning.

Therefore, taking urban villages as the largest power supply area, based on the actual situation of the government basic grid, urban-rural planning, power source location, load distribution, a grid-based method is adopted to reasonably partition the low-voltage distribution network of urban villages, and load forecasting is carried out for the power supply partition. Based on the load forecasting, urban village distribution network planning is carried out, and a construction plan for urban village power facilities is proposed. The distribution substations within the power grid can adjust the load to each other, and a low-voltage connection scheme is formed through the end lines externally to meet the development needs of urban village electricity load and improve power supply reliability and investment accuracy. The specific methods and ideas for grid partitioning of low-voltage power supply grids in urban villages are as follows:

- 1) The power grid zoning of urban villages should meet the relevant requirements for power supply area classification and low-voltage distribution network in the planning and design technical guidelines.
- 2) The grid division of power supply in urban villages should fully utilize digital means, comprehensively and accurately sort out the information of urban villages and districts, achieve low-voltage transparency, and lay the foundation for the division work.
- 3) The division of the power grid in urban villages should be combined with administrative divisions, geographical conditions, urban and rural planning, and other factors, with prominent geographical features such as rivers, mountains, and main roads as the boundaries of the power grid.
- 4) The power grid of urban villages should be combined with the regional load situation, development level, and development positioning to divide distribution transformer substations with similar development levels into the same power grid while ensuring that there is no duplication or leakage between grids. By forming a planning unit through the power grid, the coupling degree and planning complexity of the power supply area can be effectively reduced.
- 5) Under normal operation, the power grid in urban villages operates relatively independently in each zone, and load adjustment and transfer can be carried out within the power grid. In special circumstances, the power grids should have a certain degree of mutual support ability.

Here, a preliminary model for evaluating the grid of urban villages is established in total 19 typical key indicators are selected from five dimensions: power supply capacity, equipment operation, power quality, equipment level, and intelligence level, as shown in Table 1.

Table 1. Indicators for low-voltage power distribution systems

Number	Primary Indicator	Secondary Indicator	Measurements
1	Power Supply Capacity	Daily Overload Tap Changer Ratio (Average Current)	%
2		Overload Limit Exceeding Transformer Ratio (Average Current)	%
3		Average Load Rate on Peak Load Days	%
4	Equipment Operation	Daily Overload Tap Changer Ratio (Maximum Current)	%
5		Branch Overload Tap Changer Ratio (Maximum Current)	%
6		Average Outage Times per Unit	times
7		Total Number of Grid Outages	times
8		Average Number of Power Supply Reliability-Related Claims per Unit	times
9		Total Number of Grid Power Supply Reliability-Related Claims	times
10	Power Quality	General Three-Phase Imbalance Transformer Ratio	%
11		Severe Three-Phase Imbalance Transformer Ratio	%
12		Severe Low Voltage Transformer Ratio	%
13		Proactive Severe Low Voltage Transformer Ratio	%
14		General Low Voltage Transformer Ratio	%
15	Equipment Level	High Loss Transformer Ratio	%
16		Aging Transformer Ratio	%
17		Aging Low-Voltage Cabinet Ratio	%
18	Level of Intelligence	Smart Grid Coverage Rate(including 2.0 platform)	%
19		Low-Voltage Interconnection Rate	%

3.2 Proposed Partitioning Solution

Based on clarifying the grid index system, combined with the requirement of combining geographical distribution, urban and rural land planning and other factors in the grid division criteria, a method based on a genetic algorithm is proposed to splice the government basic grid. By modeling the power grid division problem as an optimization problem

and searching for the optimal solution through the heuristic algorithm, a locally optimal grid splicing scheme can be obtained. At the same time, the use of graph technology to process government grids effectively ensures that the solutions obtained by heuristic algorithms satisfy adjacency constraints. Based on the above method, the integration of power grid and government grid planning can be achieved, optimizing the existing low-voltage power grid division scheme for urban villages.

Firstly, using graph technology, the grid is modeled as a graph structure. Each government grid can serve as a node in the graph, and the adjacent relationships between different government grids can serve as edges in the graph.

The spatial correlation between nodes can be represented by a graph $G = (V, E)$, which $V = \{v_1, v_2, \dots, v_n\}$ represents the set of all nodes, and the characteristics of each node can be represented as a vector: [center longitude, center latitude, load characteristics, number of transformers in the grid, land use properties, upper-level substation]. To quantitatively describe the graph structure, an adjacency matrix $A_{ij} \in R^{N \times N}$ is introduced. Here, the following can be obtained:

$$A_{ij} = \begin{cases} 1, & \text{Node } i \text{ is connected to Node } j \\ 0, & \text{else} \end{cases} \quad (1)$$

After modeling the government grid as a graph structure, the problem of government grid stitching can be modeled as an optimization problem. Introducing the internal similarity index k_s , the variance of characteristic variables and the non-uniformity of categorical variables of each government grid within the power grid are normalized to characterize the similarity between government grids within the unified power grid. To obtain the internal similarity of each power grid, this work takes the average internal similarity of each power grid as the internal similarity of the entire region and uses this to measure the quality of the partitioning scheme. The smaller the internal similarity index, the better the partitioning scheme. The optimization objective of minimizing the region can be represented by Eq. (2).

$$\left\{ \begin{array}{l} \min k_s \\ k_s = \frac{1}{U \sum_{k=1}^{m+n} w_k} \left[\sum_{u=1}^U \left(\sum_{k=1}^m w_k M_{u,k}^s + \sum_{k=m+1}^{m+n} w_k N_{u,k}^s \right) \right] \\ M_{u,k}^s = \text{var} \left(\frac{2v_{g,u,k}}{v_{\max,u,k} - v_{\min,u,k}} \right) \\ N_{u,k}^s = \frac{T_{u,k} - 1}{G_u - 1} \end{array} \right. \quad (2)$$

The constraints of this optimization problem are as follows:

- 1) Neighbor relation constraint: After generating the scheme, extract subgraphs based on the selected nodes in the grid to determine if they are connected; If not connected, punishment will be imposed.
- 2) Non-crossing constraint: For large rivers, when establishing the adjacency matrix, it should be assumed that the grids on both sides cannot be connected and the adjacency relationship needs to be corrected.
- 3) Constraint on the number of transformers: The number of transformers in each power grid should be controlled within 4–6. If the divided power grid exceeds this range, punishment will be given according to the degree of violation.

After modeling the government grid stitching problem as an optimization problem, it can be solved through various methods. Due to the numerous penalties and conditional judgments involved in the constraints of the government grid stitching problem, heuristic methods are more suitable for solving this problem. In this work, the genetic algorithm is adopted to solve the optimization problem. It searches through populations and can simultaneously explore multiple regions in the solution space, thereby increasing the probability of finding the global optimal solution. This feature makes genetic algorithms particularly effective in handling complex, multimodal, high-dimensional optimization problems, as it can avoid the shortcomings of traditional algorithms that are prone to getting stuck in local optima.

3.3 Electricity Demand Analysis and Characterization

The load data of the substation and the electricity consumption data of users are the main parts that constitute the electricity consumption data of the power system. These two parts of the data describe the main electricity consumption characteristics of the substation and users. This study selected single-day 96 o'clock substation load data and transformer downstream user electricity data as the research objects for load characteristic analysis. The load data of the substation area is recorded daily and can be used to characterize the electricity consumption characteristics of different substations at the daily level; The 96-point current data is collected every 15 min, which can describe the load changes of the transformer within 24 h a day and can be used to characterize the load characteristics of the transformer load on a smaller time scale. This section will combine load data and user electricity consumption data to explore electricity consumption data and load characteristics in the substation area.

The process of demand analysis mainly includes the following steps.

- (1) Data preprocessing: Firstly, in the case of missing load data at a certain moment in the substation area, linear interpolation is performed to fill in the corresponding load of the substation area. If the data in a certain area is missing more than 40% on a certain day, it is considered that there is too much data missing for that day and it will be discarded; If more than 3 consecutive data are missing (i.e. continuous one-hour data missing), the data is considered invalid and discarded.
- (2) Typical daily load curve extraction: By plotting the annual/monthly load curve of the substation area, the K-means clustering method is used to extract the daily load curve of the substation area as the typical daily load curve of the substation area.
- (3) Load indicators: To characterize the load characteristics of the substation area, some load indicators need to be selected. Based on typical daily load curves, corresponding daily/monthly load indicators can be selected, such as daily maximum load, daily minimum load, monthly maximum load, monthly minimum load, monthly load peak valley difference, etc. Use these indicators to describe the load characteristics of the substation area.

In this work, the release coefficient method is a method used to analyze and characterize the load characteristics of substations, especially in the power system for evaluating and predicting load changes at different periods. This method describes the characteristics of load changes over time by releasing the concept of coefficients, providing an

important basis for the planning and operation of power systems. The main idea behind the release coefficient method is to calculate and utilize the release coefficient to describe the load characteristics. The release coefficient refers to the ratio of the load during a specific period to the load during the reference period (usually the peak load period). By analyzing the release coefficient, we can understand the relative changes in load over different periods.

The release coefficient method is essentially a multiple linear regression model, which is used to describe the linear relationship between the dependent variable and multiple independent variables. Its mathematical expression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (3)$$

Here, Y is the dependent variable, X_1, X_2, \dots, X_n is the independent variable, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ is the regression coefficient, and ϵ is the error term.

The release coefficient method predicts and analyzes load characteristics by constructing a linear relationship between the load data of each time period and the benchmark time period load. In the release coefficient method, assuming a benchmark load P_{base} and considering the proportion of industrial, agricultural, commercial, and other users in the substation area, it can be transformed into a multiple linear regression problem:

$$P_i = \beta_0 + \beta_1 P_{base,i-1} + \beta_2 p_{agr} + \cdots + \beta_n p_{res} + \epsilon \quad (4)$$

Among them, P_i is the load of the time period, $P_{base,i-1}$ represents the benchmark load of the previous year, p_{agr} represents the proportion of agriculture in the area, $\beta_1, \beta_2, \dots, \beta_n$ represents the multiple linear regression coefficient, that is, the release coefficient. Through regression analysis, the impact of the proportion of different types of users in different areas on their load can be obtained.

4 Experiments and Numerical Results

4.1 Experimental Settings

The experiments are carried out based on the dataset obtained from the real distribution network consisting of a range of power outage events in 13495 distribution transformers' daily operation measurement records. The time series of transformer operation monitoring data are sampled at a frequency of 15 min. The government grid data and transformer data used for grid partitioning are provided by the power supply company, while the road network data is sourced from public Gaode data. In the analytical study of load characteristics, 25 station areas in Huangpu Village, Guangzhou, are used as the research object to study the relationship between the monthly maximum load of the station area and the proportion of different user types (residential, commercial, industrial) in the station area, and the monthly average load of the station area is used as the base load, and the release coefficient method is used to carve out the change of the monthly maximum load of each station area in each month; the data of 2021 and 2022 are used as the training set, and the data of 2022 and 2023 monthly data are used as validation. All mentioned algorithms are implemented in Python.

4.2 Power Distribution System Partitioning Performance

To verify the effectiveness of the algorithm, this study selected three urban villages of varying sizes from a province in southern China as test cases. Among them, Urban Village A is small-sized, Urban Village B is medium-sized, and Urban Village C is large-sized. Several government grids were stitched to form the power supply grids, with the grid partitioning results for each urban village shown in Fig. 1. In these figures, the red lines denote the boundaries of the urban village, and the polygons represent the government grid.

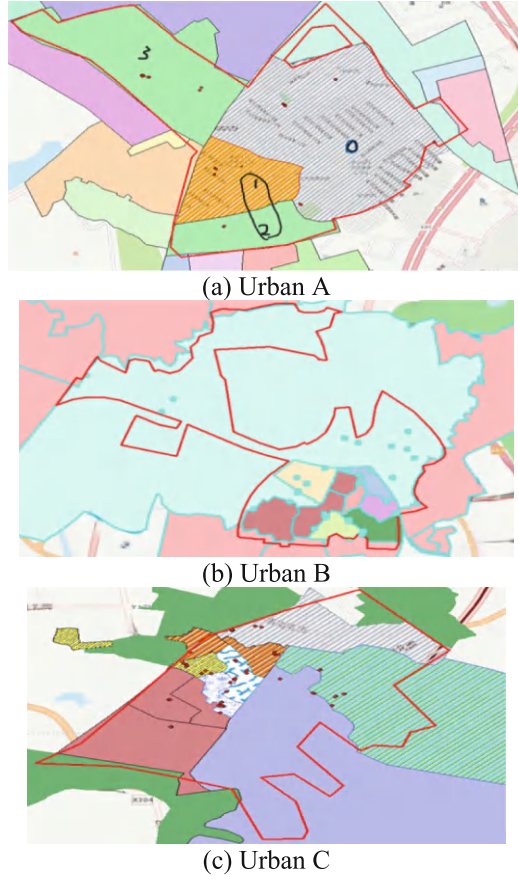


Fig. 1. The grid partitioning results for each urban village

From these figures, it can be observed that the proposed method can guarantee the interconnection of government grids without crossing roads or rivers. To further assess and quantify the results of the grid partitioning, Table 2 has been introduced.

Form Table 2, It is evident that the proposed method strictly adheres to constraints (1) and (2), thereby ensuring the usability of the partitioning results. Additionally, this

Table 2. The grid partitioning results

Urban Village	Internal Similarity	Number of transformers			Number of violated (1) (2) constraints
		Avg	max	Min	
A	0.083	7.333	9	5	0
B	0.0625	6.375	19	3	0
C	0.0549	6	9	3	0

approach effectively guarantees that the grids are geographically proximate and exhibit similar load profiles, all while keeping the number of transformers within a specified range.

4.3 Electricity Demand Characterization Performance

Based on the release coefficient method to portray the relationship between the monthly maximum load of the station area and the percentage of industrial, commercial, residential, and other types of users in the station area, the results of the study are as follows. Summer and winter load fluctuations are selected as typical scenarios, and Fig. 2(a) shows the predicted load fluctuations of 25 station areas in summer (July) and Fig. 2(b) shows the predicted load fluctuations of 25 station areas in winter (January). The blue curve is the real data and the yellow curve is the formula forecast data. The forecast deviation is 11.4% in July and 15.1% in January.

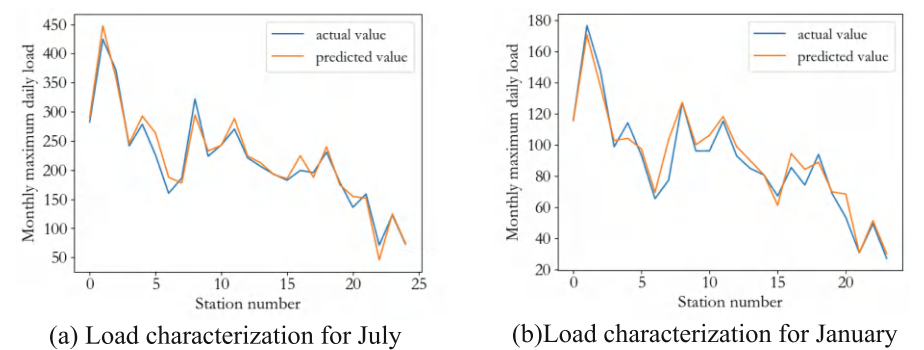


Fig. 2. Characterization Performance in typical scenarios: (a) load characterisation for July; (b) Load characterisation for January.

The study was carried out on the load for the whole year 2023 and the following table shows the forecast errors by month Table 3.

From the above table, it can be seen that the prediction error for the whole year is controlled within 20%, with an accuracy rate of more than 80%, which shows that the

Table 3. Prediction error

Month	Error
January	15.1%
February	18.2%
March	4.4%
April	19.1%
May	12.6%
June	13.5%
July	11.4%
August	18.5%
September	16.7%
October	10.3%
November	5.2%
December	12.1%

study based on the release coefficient method can better portray the relationship between the load changes of station and the percentage of the type of subscribers.

4.4 Discussions

Through research, it has been found that urban low-voltage distribution networks are crucial for ensuring the safety of urban power supply. At the same time, there are a series of challenges in efficiently ensuring the safe and stable operation of urban power grids. The following observations and findings can be observed to improve the performance of urban power distribution systems:

- (1) Strengthen the coordination and connection between urban planning and power grid planning, ensuring that power grid construction is carried out synchronously with urban development. At the same time, we will strengthen forward-looking research on power grid construction and improve scientific and rational planning.
- (2) Increase investment in the construction and maintenance of the distribution network to ensure the smooth progress of equipment updates, renovations, and daily maintenance work. At the same time, actively introducing new technologies, processes, and materials to improve the intelligence level and operational efficiency of the power grid.
- (3) Through scientific planning and rational layout, optimize the distribution of electricity load and reduce the operating pressure of the power grid. At the same time, strengthens demand side management, and guides the customers to use electricity reasonably and save electricity.
- (4) Establish a sound distribution network management system, and clarify management responsibilities and processes. Strengthen the training and education of management personnel to improve their professional skills and management level.

Finally, it is required to further strengthen the monitoring and evaluation of the operation of the power grid, and promptly identify and handle problems. Promote intelligent construction: Accelerate the pace of smart grid construction and improve the intelligence level of distribution networks. By introducing smart devices such as smart meters and smart switches, as well as advanced technologies such as big data analysis and cloud computing, intelligent management and optimized scheduling of the power grid can be achieved.

5 Conclusive Remarks

This work exploited the network partitioning and demand characterization method for management of low voltage power distribution systems. The work firstly investigated the evaluation and application strategies of power grid in urban villages, and constructed a grid indicator system from five dimensions: power supply capacity, equipment operation, power quality, equipment level, and intelligence level. In addition, the automatic partitioning method of power professional grid based on government grid is studied and a genetic algorithm-based method is developed. Further, this work constructs a substation and user load characteristic analysis model. The proposed solution is validated through a case study and the numerical results confirmed its effectiveness.

For future work, with the development of computer, communication, and network technologies, urban distribution networks will gradually move towards intelligence, achieving automation, informatization, and intelligent management of distribution networks. Promote the development and application of green energy, and facilitate the green transformation of urban distribution networks. Strengthen the interaction and communication between power supply enterprises and users, and improve the personalized and differentiated level of power supply services. In short, urban distribution network management is an important guarantee for ensuring the safety, reliability, and economic operation of urban power supply. By implementing measures such as rational planning, scientific management, and high-quality services, the operational efficiency and management level of urban distribution networks can be continuously improved, providing strong support for the sustainable development of cities.

Acknowledgment. This work is supported by the China Southern Power Grid Science and Technology project “Research and application of key technologies for low voltage situation awareness in production command mode” (030100KK52230003/GDKJXM20220007).

References

1. Fakihi, S., Mabrouk, M.T., Batton-Hubert, M., Lacarrière, B.: Impact of uncertainties in power demand estimation on the optimal design of renewable energy sources and storage systems. In: 2022 IEEE 10th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, pp. 68–73 (2022)
2. Haghgoo, R., Kojury-Naftchali, M., Fereidunian, A.: customers characterization by seasonality detection in residential electricity consumptions data, based on high-day frequency identification. In: 2023 13th Smart Grid Conference (SGC), Tehran, Islamic Republic of Iran, pp. 1–6 (2023)

3. Bignucolo, F., et al.: Characterization of residential users' behaviour and influence on distribution network planning. In: 2020 55th International Universities Power Engineering Conference (UPEC), Turin, Italy, pp. 1–6 (2020)
4. Weihong, H., Lu, F., Rong, S., Xincun, Z., Fengyuan, N.: Research on customer electricity behaviour based on K-means clustering algorithm. In: 2023 China Automation Congress (CAC), Chongqing, China, pp. 6517–6522 (2023)
5. Al Wardi, M.S.A., Jayaweera, D.: Demand characterization and management in a smart city. In: 2016 Eighteenth International Middle East Power Systems Conference (MEPCON), Cairo, Egypt, pp. 392–399 (2016)
6. Rout, B., Saraswat, G., Natarajan, B.: Efficient network partitioning: application for decentralized state estimation in power distribution grids. In: 2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, pp. 1–5 (2023)
7. Li, W., Zuo, X., Ju, L., Wang, G., Qiu, Y., Nian, H.: Research on partitioning real-time reactive power optimization method for distribution network with renewable energy. In: 2022 5th International Conference on Power and Energy Applications (ICPEA), Guangzhou, China, pp. 360–364 (2022)
8. Zhou, S., Liu, H., Guo, Z., Wang, J.: Parallel load restoration of distribution network via a shortest-path-based grid partitioning approach. In: 2022 IEEE 5th International Electrical and Energy Conference (CIEEC), Nanjing, China, pp. 1534–1539 (2022)
9. Gu, H., Chu, X., Liu, Y.: Partitioning active distribution networks by using spectral clustering. In: 2020 IEEE Sustainable Power and Energy Conference (iSPEC), Chengdu, China, pp. 510–515 (2020)
10. Lin, J., Lin, Y., Liu, W., Cheng, H., Song, Q., Lin, T.: Research on dynamic island partition of distribution network with distributed generation. In: 2021 China International Conference on Electricity Distribution (CICED), Shanghai, China, pp. 523–527 (2021)



Automation and Data Aggregation Algorithm for Low Code Development in Power Grid Mobile Report Generation

Wenting Wei^(✉) and Jie Zhang

Internet Office, State Grid Ningxia Electric Power Co., Ltd., Wuzhong Power Supply Company,
Wuzhong 751300, Ningxia, China
weiwenting@wz.nx.sgcc.com.cn

Abstract. Data management and report generation in power grid industry have extremely high requirements for accuracy and real-time. Traditional technology is not only time-consuming but also error-prone when dealing with a large amount of data. With its automation and flexibility, the low-code development platform provides new possibilities for power grid report generation. This paper introduces the low-code platform, focusing on the functions and characteristics of the low-code development platform. Secondly, functional modules are designed, including data collection, business process management, task management, automatic scheduling and asset management. Then the K-Means clustering algorithm is combined to improve the efficiency and accuracy of data aggregation. Finally, the automation of report generation process is realized. Through the experimental test, the report generation cycle of low code development is significantly lower than that of traditional technology, the average aggregation time is only 484.5 ms, and the error rate is between 0.031 and 0.069%, which is much lower than that of traditional technology. These data show that low-code development has higher accuracy and efficiency in the process of data aggregation and report generation.

Keywords: Low Code Development · Power Grid Movement Report · K-Means Clustering Algorithm · Polymerization Time

1 Introduction

The data management of power grid industry involves sensor data, equipment status, power consumption and other information. These data are not only huge, but also frequently updated, so traditional data processing methods can not meet the requirements of real-time and accuracy.

The purpose of this paper is to discuss the application of low-code development platform in the generation of power grid mobile reports. Through the experimental verification, this paper shows the obvious advantages of low-code development in improving the automation of report generation, the efficiency of data aggregation and the security and reliability of the system. The research in this paper not only provides an efficient and

reliable report generation tool for power grid enterprises, but also provides a reference for the application of low-code development platform in other industries.

Firstly, this paper introduces the background of low-code development platform and the requirements of power grid mobile report generation. Then, the selection and configuration of low-code development platform, as well as the design and implementation of functional modules are expounded. Then it introduces the application of K-Means clustering algorithm. The report generation cycle, data aggregation accuracy and system security of the low-code development platform are also tested and evaluated. Finally, the application effect of low code development in power grid mobile report generation is summarized, and the future research direction is prospected.

2 Related Work

In today's fast-moving information technology world, generating and managing production reports efficiently and accurately has become a key challenge for businesses and organizations. Peng Cheng linked multiple database fields and added logical relationship constraints to associate and display data from multiple source tables in the same unit. Finally, he verified the effectiveness of generating multidimensional reports through examples, providing technical support for flexible mining and display of data content [1]. Liang Xueqing proposed a power production report generation system design based on eXtensible Markup Language, as it takes a long time to generate power production reports using traditional systems. Experimental results had shown that the design system had a shorter time to generate electricity production reports and had good feasibility and applicability [2]. Zhu Mingxing developed an automatic production report generation and sending system using Python language to address the time-consuming and labor-intensive nature of traditional production report production, as well as the tendency for data errors. This system achieves automatic generation and sending of production reports, reducing workload and improving work efficiency, providing reference for the automation of other types of report production [3]. By integrating the existing system data, Chen Weihang established a unified template for the three business reports of production, safety and facilities, automatically captured the data available in the system, reduced the workload of report entry, realized the rapid generation and real-time sharing of daily reports, and improved paperless office [4]. Sheng Huanyu had designed an intelligent reporting system - Xiangjiaba Power Station Management Center. The system realizes the function of automatically generating work reports such as operation analysis, maintenance analysis, and earthquake special analysis for Xiangjiaba Power Station, which greatly improves the accuracy, integration, intelligence level, and actual efficiency of the power station production report system data [5].

Moreover, Di Ruscio D compared and contrasted low code and model driven methods, identified their differences and similarities, and analyzed their advantages and disadvantages [6]. Phalake V S analyzed existing low code platform technologies and evaluated their advantages and limitations in application development [7]. Malik H critically examined the statement that no code and low code platformed promise to enable individuals with minimal coding experience to build applications, exploring the potential advantages and disadvantages of low code development [8]. Hurlburt G F summarized

the existing knowledge of low code development from multiple perspectives, including definitions, tools used, application development lifecycle, application domains, potential benefits, challenges, and related development and delivery principles [9]. Umaroh S used a low code platform to develop applications to control the capacity of the mosque and avoid physical contact during prayer [10]. The flexibility and adaptability of existing automated report generation systems still need to be further improved in the face of changing business requirements and environmental changes. In order to conscientiously implement the spirit of the “Two Sessions” and the 2023 Digital Work Conference of State Grid Ningxia Electric Power Co., Ltd., deepen the implementation of digital transformation, promote the landing of new technological achievements in business processes, and standardize the workflow of various digital professional business orders, enhance the lean level of digital business management, provide convenient office services for employees, improve work quality and efficiency, and effectively support the high-quality development of the company and the power grid, by analyzing the capabilities and limitations of low code development platforms, combined with data aggregation technology, this article aims to propose a new report generation solution to achieve more efficient and intelligent report generation and management in the field of power grid management.

3 Method

3.1 Low Code Platform

Power grid industry has specific requirements for data processing and report generation, and low-code development provides rich functional modules and intuitive visual development tools, which is very suitable for power grid industry applications.

The form designer developed with low code supports a variety of field types and advanced form logic, and flexibly designs forms in the data acquisition stage to meet different data input requirements. Moreover, low-code development provides powerful data processing components and script support. Through these tools, complex data processing logic is set, and data cleaning, conversion and aggregation are carried out. Data display is the core of report generation, while the dashboard and visual components developed with low code design an intuitive report display interface. Using charts, indicator cards and maps to show key indicators and trends. These reports are not only viewed on the desktop, but also accessed through mobile devices to ensure that grid workers get the information they need.

Security is an important consideration in power grid data management. Low code development supports data encryption and access control to ensure data security. In addition, the platform also supports secondary development and local privatization deployment, which is convenient for enterprises to expand system functions according to their own needs.

3.2 Functional Module Design

Through low code development, power grid enterprises design and implement specific functional modules, thus improving the automation level of data management and business processes.

Using the business process management module developed with low code, power grid enterprises have transformed the traditional approval process into online circulation, realizing automation. This module supports file uploading, so that all relevant documents can be easily attached to the approval process. Approvers can view the contents of files in the system through the online preview function, while encryption and decryption operations ensure the security of sensitive data and documents.

The task management module allows users to create task descriptions, including task requirements, expected results, deadlines and priorities. The task assignment function enables managers to assign tasks to team members, while the task monitoring function provides real-time updates, showing the progress and status of tasks.

The automatic scheduling module realizes efficient scheduling through preset rules and conditions. This module automatically arranges work shifts according to employees' skills, workload and availability. Considering the power demand of the power grid industry during peak hours, the automatic scheduling system flexibly adjusts shifts, reduces the workload of manual scheduling, and ensures the fairness and rationality of scheduling.

The realization of asset management module automatically imports terminal ledger information through the data interface developed with low code, which improves the automation level of asset management. It tracks and manages all power grid assets, including equipment, tools, and vehicles, automatically updates asset status such as maintenance, replacement, and scrapping, ensuring the accuracy and timeliness of asset management. At the same time, it provides historical records of asset usage and maintenance to help managers make more informed decisions.

3.3 Data Aggregation Algorithm Development

In the power grid mobile report generation, low code development improves the efficiency and accuracy of data aggregation. By developing efficient data aggregation algorithms, power grid enterprises integrate data from different business systems to ensure data consistency and availability. The specific data are shown in Table 1 [11, 12]:

Table 1 shows the key operating parameters of different business systems in the grid, including power consumption, voltage, current, equipment status, peak load, and maintenance schedule. With these data, grid companies comprehensively monitor and analyze the operation of the power system.

The K-Means clustering algorithm divides the data points into k clusters so that the data points within the same cluster are as similar as possible, and for each data point x_i , its distance from the center of each cluster is computed and it is assigned to the nearest cluster center:

$$\text{Assign } x_i \rightarrow \operatorname{argmin}_z ||x_i - \mu_z||^2 \quad (1)$$

μ_z is the center of the z th cluster.

The Means clustering algorithm minimizes the objective function J through an iterative process:

$$J = \sum_{z=1}^K \sum_{x_i \in S_z} ||x_i - \mu_z||^2 \quad (2)$$

Table 1. Business system data

System	Electricity Consumption (kWh)	Voltage (V)	Current (A)	Equipment Status	Peak Load (MW)	Maintenance Schedule
Transmission Control	1200	220	5.5	Normal	500	Q2 2024
Distribution Management	1100	230	6.0	Alert	450	Q3 2024
Load Forecasting	1000	210	4.8	Fault	400	Q1 2024
Fault Detection	1300	215	5.2	Normal	550	Q4 2024
Power Dispatching	1400	240	6.5	Normal	600	Q2 2025
Energy Trading	1500	250	7.0	Alert	650	Q1 2025
Billing System	1600	260	7.5	Normal	700	Q3 2025

where S_z is the set of data points in the z th cluster. The implementation of the algorithm consists of initializing the cluster centers, assigning data points to the nearest cluster centers, updating the cluster centers, and iterating these steps until the cluster centers no longer change significantly.

On the low-code development platform, the data is collected and input through the form designer, and the data table function is used to clean and preprocess the data, which ensures that the data is accurate and consistent before entering the clustering algorithm. Next, the script support function provided by the platform is used to write the implementation code of K-Means clustering, including defining data points, initializing cluster center, calculating the distance from data points to cluster center, updating the location of cluster center, etc. [13, 14].

According to the use frequency and fault history of equipment, in the analysis of equipment maintenance, the equipment is clustered, the maintenance plan is optimized, and the operation efficiency of equipment is improved [15]. In addition, by analyzing the equipment status data, potential failure modes can be identified, and intervention can be made in advance to reduce the risk of failure.

3.4 Automation and Optimization

In the power grid mobile report generation, through low-code development, power grid enterprises design automatic workflows, which automatically collect and process data from different power grid business systems, such as substation monitoring, distribution network management, load forecasting, fault detection, etc., and integrate these data into a unified report. The realization of automatic process is based on preset templates and rules, which ensures the consistency and standardization of report generation and

reduces the possibility of human error. In addition, automatic report generation also allows automatic execution according to the set schedule or specific trigger conditions, which further improves the timeliness of report generation.

In addition, system optimization ensures the long-term stable operation and continuous improvement of the power grid mobile report generation system. By collecting user feedback and analyzing system operation data, power grid enterprises identify the improvement points of system performance and user experience. These optimization measures include improving user interface to improve usability, enhancing data processing ability to support more complex analysis, and optimizing report presentation to provide more intuitive information presentation. Continuous system optimization not only improves the reliability of the system, but also enhances the satisfaction of users, ensuring that the system can adapt to the changing business needs and market environment.

3.5 Security and Reliability Guarantee

In the power grid mobile report generation, the application of low code development not only improves the efficiency of automation and data aggregation, but also provides a solid guarantee for security and reliability.

Through low code development, power grid enterprises implement strict information security measures, which meet the information security requirements of State Grid Corporation of China. This includes encrypting sensitive data, ensuring the security of data during transmission and storage, and preventing data leakage. At the same time, the platform supports the security audit function, records and monitors the operation behavior of all users, and timely discovers and responds to potential security threats.

Secondly, the low-code development platform has designed high concurrent access support to ensure that the system still runs stably when a large number of users access the system at the same time. This not only improves the user experience, but also ensures the timely generation and distribution of key reports. In addition, the stability and fault tolerance of the platform enable the system to continuously provide services in the face of various abnormal situations that may occur in the power grid industry, and meet the demand of the power grid industry for high reliability.

4 Results and Discussion

4.1 Implementation Effect

In the power grid mobile report generation, by using the low-code development platform, power grid enterprises build a highly automated report generation process, which automatically extracts the required data from various data sources in the power grid. The data in these data sources are automatically extracted and processed by predefined logic to generate standardized reports, which reduces human intervention and thus reduces data inconsistency caused by human operation errors.

In power grid data aggregation, K-Means algorithm can gather data points with similar characteristics together, thus helping to identify patterns and trends in data. The

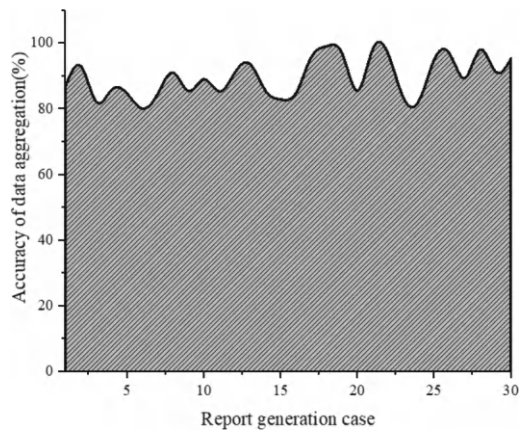


Fig. 1. Accuracy of Data Aggregation

data aggregation accuracy of K-Means algorithm in different mobile report generation cases is shown in Fig. 1:

By analyzing the test data in Fig. 1, it can be seen that the K-Means algorithm has obvious advantages in the mobile generation of power grid reports, and its data aggregation accuracy has remained above 80.1%, and it has reached 99.1% in the 18th group of report generation cases. Behind this figure, it is the algorithm’s profound insight and accurate capture of the internal structure of power grid data. By iteratively calculating the cluster center, K-Means algorithm can effectively distinguish different types of data points, and can maintain a high degree of accuracy even in the face of complex and changeable power grid data.

Moreover, this paper also tests the data aggregation time of the algorithm to measure its real-time data processing ability, and the results are shown in Fig. 2:

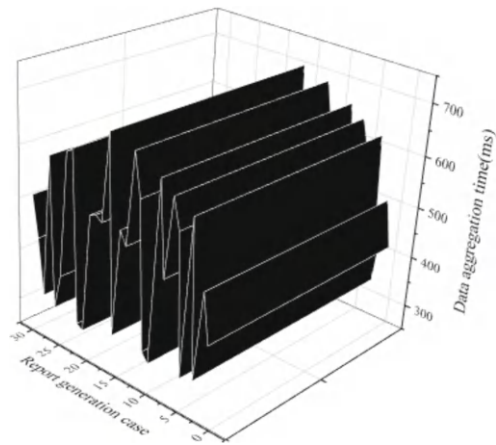


Fig. 2. Data aggregation time

This paper tested the aggregation time of the algorithm in report generation and summarized its data in Fig. 2. This paper further analyzes its data aggregation time and finds that among the 30 report generation cases, the aggregation time fluctuates relatively small, ranging from 304 to 694 ms, with an average of 484.5 ms. This indicates that the algorithm has good consistency and robustness on different datasets. This consistency means that the algorithm provides reliable and timely report data in any situation, thereby supporting the stable operation and effective management of the power grid.

4.2 Performance Evaluation

Based on low code development, this paper realizes the generation of power grid mobile report. In terms of report generation cycle, it allows users to quickly build report templates, configure data sources and aggregation logic, and realize automatic report generation without programming. This efficient development method shortens the whole process time from data collection to report presentation, and enables power grid enterprises to grasp the operation status and make decisions in a timely manner. In contrast, in the traditional method, developers write a lot of code and design complex data processing flow, which leads to a long report generation cycle. The specific comparison results are shown in Fig. 3:

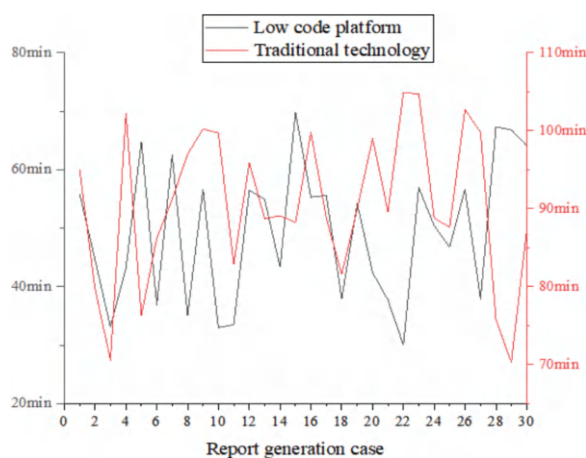


Fig. 3. Report generation cycle

As shown in Fig. 3, the report generation cycle of low code development is generally lower than that of power grid mobile report under the traditional technology. In the first group of cases, the report generation cycle of low-code development is 55.9 min, but it reaches 95 min under the traditional technology, which exceeds the report generation cycle of low-code development by 39.1 min. This great improvement in efficiency is mainly due to the automatic data processing and report generation capabilities of the low-code development platform. In low-code development, data extraction, cleaning, aggregation and report generation are automatically completed through preset processes, which greatly reduces manual intervention and waiting time. In contrast, the

traditional technology manually extracts and processes data, which is cumbersome and time-consuming.

This paper also compares the error rate of report generation, and the result is shown in Fig. 4:

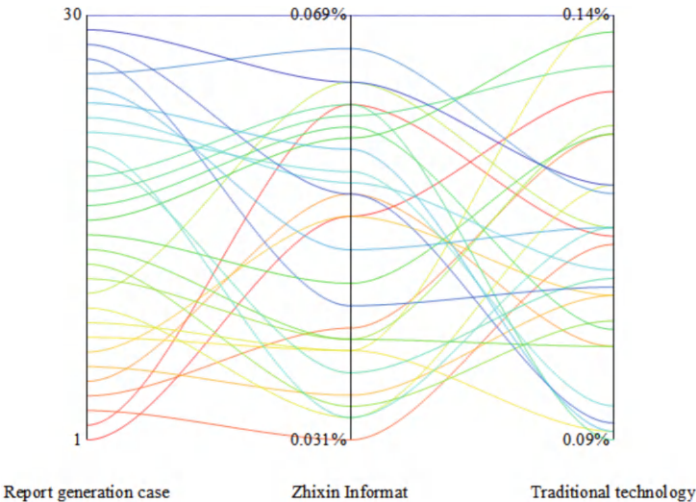


Fig. 4. Error rate of report generation

The data in Fig. 4 clearly shows that the error rate of report generation with low code development is between 0.031 and 0.069%, while that with traditional technology is between 0.09 and 0.14%. This low error rate means that the errors introduced by low code development in the process of data aggregation and report generation are minimal, thus providing more reliable and accurate decision support for power grid enterprises.

4.3 Discussion

Compared with the traditional technology, in the power grid mobile report generation system, the application of low code development improves the automation of report generation and the efficiency of data aggregation. Low code development shortens the report generation cycle from 95 to 55.9 min, which greatly improves the real-time and accuracy of data. This technological progress not only reduces the time cost, but also enables power grid managers to gain insight into key operating indicators and abnormal situations more quickly.

Nevertheless, this study also encountered some technical challenges. Data consistency is the key issue. Therefore, the data cleaning and preprocessing module is introduced in this study to automatically detect and correct data errors, and the data consistency is ensured through the data verification mechanism. User experience and system integration are also the focus of attention. This study provides more customization options and templates, allows users to adjust the interface layout and functions according to their needs, and supports the development of plug-ins and extensions to increase

the customization of the system. At the same time, by adopting open standards and API interfaces, the interoperability of the system is ensured, and sufficient system integration tests are carried out to ensure seamless cooperation among various components.

5 Conclusion

In this paper, the low-code development platform has shown remarkable advantages in the automation of report generation and the accuracy and efficiency of data aggregation. The low-code development platform enables power grid enterprises to quickly build a report generation process and realize the automation of the whole process from data collection to report display through its intuitive visual development tools and rich functional modules. In the aspect of data consistency, by introducing data cleaning and preprocessing module, data errors can be automatically detected and corrected, and data consistency in the aggregation process can be ensured through data verification mechanism, which improves the accuracy of data. In addition, the enhancement of real-time data processing capability enables power grid enterprises to quickly respond to various situations in power grid operation and optimize decision-making. Nevertheless, there are some limitations in this study. The main focus is that the low-code development platform fails to cover the potential advantages of other low-code development platforms. At the same time, the experimental tests are mainly based on specific power grid data sets, which show different performances in different data sets or power grid environments. The scalability and long-term maintainability of the system need to be verified in a wider range of application scenarios. The future low-code development platform has a broad application prospect in the generation of power grid mobile reports. With the continuous progress of technology, the function and performance of low-code platform will be further improved, providing more efficient and intelligent report generation tools for power grid enterprises. Future research can focus on multi-platform comparison, compare the application effects of different low-code development platforms in power grid mobile report generation, and provide more comprehensive choices for power grid enterprises.

References

1. Cheng, P.: Custom multidimensional report generation based on data mining. *Comput. Knowl. Technol.* **19**(4), 69–71 (2023)
2. Liang, X., Du, S., Liu, L.: Design of an XML based power production report generation system. *Autom. Appl.* **64**(1), 141–143 (2023)
3. Zhu, M., Shi, J., Hu, Y.: Development and application of an automatic production report generation and sending system based on Python. *Hongshuihe* **41**(3), 139–142 (2022)
4. Chen, W.: Application of intelligent reporting system in a certain oilfield in Bohai Sea. *New Ind.* **12**(1), 6–7 (2022)
5. Sheng, H., Lai, J.: Design and application of intelligent reporting for hydropower stations based on Finereport. *People's Changjiang* **53**(S01), 201–203 (2022)
6. Di Ruscio, D., Kolovos, D., de Lara, J., Pierantonio, A., Tisi, M., Wimmer, M.: Low-code development and model-driven engineering: Two sides of the same coin? *Softw. Syst. Model.* **21**(2), 437–446 (2022). <https://doi.org/10.1007/s10270-021-00970-2>

7. Phalake, V.S., Joshi, S.D.: Optimized low code platform for application development. *Int. J. Contemp. Archit* **8**(2), 1–8 (2021)
8. Malik, H.: The rise of no-code and low-code development: Democratizing software development or creating a digital underclass? *Kashf J. Multidiscip. Res.* **1**(2), 49–57 (2024)
9. Hurlburt, G.F.: Low-code, no-code, what's under the hood? *IT Prof.* **23**(6), 4–7 (2021)
10. Umaroh, S., Putra, K.R., Barmawi, M.M.: Low-code platform for health protocols implementation in Sabilussalam Mosque during The COVID-19 pandemic. *REKA ELKOMIKA: Jurnal Pengabdian Kepada Masyarakat* **3**(2), 96–105 (2022)
11. Yousefpoor, E., Barati, H., Barati, A.: A hierarchical secure data aggregation method using the dragonfly algorithm in wireless sensor networks. *Peer-to-Peer Netw. Appl.* **14**(4), 1917–1942 (2021)
12. Pichumani, S., et al.: Ruzicka indexed regressive homomorphic ephemeral key benaloh cryptography for secure data aggregation in WSN. *J. Internet Technol.* **22**(6), 1287–1297 (2021)
13. Mohammadali, A., Haghighi, M.S.: A privacy-preserving homomorphic scheme with multiple dimensions and fault tolerance for metering data aggregation in smart grid. *IEEE Trans. Smart Grid* **12**(6), 5212–5220 (2021)
14. Joshi, P., Raghuvanshi, A.S.: A dual synchronization prediction-based data aggregation model for an event monitoring IoT network. *J. Intell. Fuzzy Syst.* **42**(4), 3445–3464 (2022)
15. Bomnale, A., More, A.: Node utilization index-based data routing and aggregation protocol for energy-efficient wireless sensor networks. *J. Supercomput.* **80**(7), 9220–9252 (2024)



Big Data Analysis and Smart Grid Security Event Monitoring and Response in the Power Internet of Things

Hongyu Ke¹(✉), Zhaoyu Zhu¹, Shuo Yang¹, Yi Tang¹, Ning Xu¹, and Xin He²

¹ State Grid Information and Communication Branch of Hubei Electric Power Co., Ltd.,
Wuhan, China

kehongyu@2980.com

² Wuhan Diameter Technology Co.,Ltd., Wuhan, China

Abstract. With the continuous development of China's smart grid and power Internet of Things, China has formed a large amount of big data. These big data have the characteristics of large quantity, diverse types, high value, and fast speed, laying a solid foundation for promoting large-scale data applications. On the power generation side, large-scale grid integration of new energy sources such as wind and solar energy has broken the traditional relative static state, making the measurement and management of electricity usage more complex. Secondly, due to the inability to store electrical energy, the safety situation in the power industry is very complex. On the power generation side, with the continuous evolution of the new generation power grid, the grid supply chain based on high elasticity and big data will gradually be replaced. This article is based on the above issues, exploring the big data analysis and smart grid security event monitoring and response in the power Internet of Things, and constructing a mathematical model through Bayesian network algorithm for power grid security event monitoring and early warning. The experimental results show that the Bayesian network model has a false alarm rate of 0.03 at low loads, which is lower than other methods.

Keywords: Grid Security · IoT Data Analysis · Bayesian Network · Power Grid Status

1 Introduction

The power Internet of Things, as an important support for the new generation of power grids, integrates new technologies such as intelligent sensing, network communication, and big data analysis, which can effectively enhance China's new energy consumption capacity, providing strong support for the development of related fields such as electric vehicle access and charging. In this context, as a new type of power system, the power Internet of Things requires real-time monitoring and diagnosis. Therefore, this article studies cutting-edge technologies such as power Internet of Things and big data analysis. The online monitoring and condition maintenance of power equipment is an interdisciplinary integration of electrical insulation, high voltage, sensing, digital

signal processing, electronics, computer, and other disciplines. Performing equipment predictive maintenance is an important means to ensure the safe and reliable operation of equipment, as well as an important supplement and innovation to traditional offline preventive testing.

In Sect. 3 “Methods”, two methods for power grid safety risk assessment are first introduced: the risk management based safety risk assessment method and the probabilistic transient stability based safety risk assessment method. Among them, the risk management based method utilizes the accident tree safety assessment model for qualitative analysis and quantitative calculation of the system, while the probabilistic transient stability based method models faulty components and solves them using probability theory. In addition, the evaluation process of event driven risk indicators and the structural learning method of Bayesian networks were also introduced. In Sect. 4 “Results and Discussion,” the process of scene generation and reduction is first described to obtain representative scenes. Subsequently, it presented the results of evaluating power grid security risks through different methods, including the accuracy of Bayesian network models and comparative analysis of other methods. Finally, this article evaluates the efficiency and performance differences of various methods by comparing their training time and false alarm rate for security alarm events.

2 Related Works

Experts have long conducted specialized research on data analysis of the Internet of Things. Alavikia Z summarized the advantages and challenges of the Internet of Things in SG (smart grid), proposed a hierarchical method to classify the applications of Internet of Things technology, and discussed future measures [1]. Bi Z introduces the Internet of Things, big data analysis, and digital manufacturing as representative technologies for data processing, using Shannon entropy to measure system complexity and emphasizing the role of electronic devices in managing system stability. He proposed a new enterprise architecture to enhance system flexibility and adaptability, emphasizing the importance of adaptability for manufacturing systems [2]. Venu D N explored the role of the Internet of Things in various fields, reveals technological challenges and opportunities, explores current challenges in IoT research, and explores potential solutions related to industrial IoT supported by 5G [3]. Akhter R explored the application of IoT based agricultural data analysis and machine learning technology to improve crop yield and quality [4]. Kumar R L explored the combination of blockchain and the Internet of Things, forming a new concept of industrial Internet of Things blockchain. He provided a detailed introduction to the relevant aspects of blockchain, including the concepts of the Internet of Things and blockchain, the challenges faced, and the proposed structural design for their integration. A survey shows that blockchain can be used to develop complex interconnected environments and achieve secure communication in decentralized networks [5]. Sharma S conducted a detailed investigation into solutions in the fields of Wireless Sensor Networks (WSN) and Internet of Things (IoT). He discussed their advantages and disadvantages and compared them based on performance metrics. These knowledge provide guidance for network architects to choose solutions suitable for specific applications [6].

Al-Khatib A W studied the relationship between IoT and big data analytics (BDA), Supply chain visibility (SCV), and operational performance (OP) in Jordan's pharmaceutical manufacturing industry by establishing a conceptual model. He used structural equation modeling for data analysis and validated various validity and hypotheses. The results indicate that the Internet of Things and BDA have a significant positive impact on SCV and OP, and there is also a significant positive relationship between SCV and OP, and SCV plays a mediating role between the Internet of Things, BDA, and OP [7]. Issa W comprehensively reviewed blockchain based federated learning methods to ensure the security of IoT systems. He introduced the application of blockchain in Federated Learning (FL), addressed IoT security issues, and outlined necessary measures to protect IoT ecosystem security and privacy [8]. Khan I H collected shared information through connected devices and controls it through identification codes, bringing disruptive innovation to the manufacturing industry. His research focuses on its potential, driving factors, and smart factory creation in Industry 4.0. This successful IoT application has improved production efficiency, reduced costs, and reduced errors, but achieving full benefits still requires sustained efforts [9]. Kishor A collected patient data through IoT sensors and combines machine learning to predict various diseases. The study used seven classification algorithms, such as decision trees and support vector machines, to successfully predict nine diseases. The performance evaluation shows that the RF classifier performs well in accuracy, sensitivity, specificity, and AUC (Area Under the Curve), providing early diagnostic support for doctors [10]. Islam M M provides a clear overview of IoT technology, including common device functionalities, architectures, and protocols. He introduced common hardware platforms such as Raspberry Pi, Arduino, and ESP8266, as well as software platforms and recently developed architectures [11]. Rahmani A M outlined the importance and security challenges of the Internet of Things in the automation century, emphasizing that cost, real-time performance, security, and errors are the main factors affecting the evaluation of IoT applications [12].

Bhoi S K proposed an IoT based environmental monitoring system that utilizes the Arduino platform and volunteer resources for remote control and access. The system records environmental information through sensors and uploads it to a network server for data analysis. He used R Language and R Studio IDE for machine learning (ML) data modeling, accurately predicting trends in temperature, humidity, carbon monoxide, and carbon dioxide levels. The experiment compared the prediction accuracy of different ML techniques in different scenarios [13]. Babangida L explored methods for human activity recognition in smart homes, emphasizing the feasibility and efficiency of IoT sensor technology. He achieved activity recognition through preprocessing, feature extraction, and classification algorithms. However, issues such as insufficient data, imbalance, and computational complexity still exist [14]. Huo R comprehensively explored the potential and challenges of applying blockchain technology to the Internet of Things, introduced the characteristics of blockchain, and how to apply it to the Internet of Things to improve service quality and promote development [15]. Shaikh F K conducted research on the deployment architecture of traditional and intelligent agriculture by optimizing management decisions to achieve more benefits, with a focus on the various processing stages of intelligent agriculture [16]. Chi Y discussed the importance of knowledge-based fault diagnosis methods in industrial IoT systems. Compared to model-based and data-driven

diagnostic methods, knowledge-based methods are more popular and can improve interoperability through ontology, providing advanced reasoning and query response for non professional users. He reviewed the latest progress in building knowledge bases through ontology and inference, and discussed the application of inductive reasoning methods in fault diagnosis [17]. Kasilingam D aimed to understand the factors that influence consumers to adopt IoT products offered in the form of services and pay monthly premiums. He analyzed the sample data using a partial least squares structural equation model and found that perceived playability, personal innovation, and convenience value were the main variables affecting attitudes and intentions. The effectiveness evaluation of predictions shows that the model has high predictive ability, which can help marketers segment the market and design business strategies, promote the dissemination and adoption of IoT services [18]. The data analysis technology of the Internet of Things faces many challenges, including data quality and security, data integration and real-time performance, algorithm processing, and system scale. To solve these problems, it is necessary to comprehensively consider factors such as data management, security, algorithm optimization, and infrastructure, in order to promote the progress and application of animal networking data analysis technology.

3 Methods

3.1 Monitoring Technology for Power Grid Security Incidents

(1) A Security Risk Assessment Method Based on Risk Management

This method requires building a model of the power system, and then conducting qualitative analysis and quantitative calculations on the system based on the model. The most representative method is the fault tree safety assessment model. The accident tree safety evaluation model is also a type of evaluation model that needs to be considered in this study. The characteristics of this method are simple and clear, which is conducive to further research on the system. This type of method has a certain degree of subjectivity. In order to evaluate the system more objectively and scientifically, evaluators are required to establish scientific indicator content and reasonable indicator weights. However, how to reasonably determine indicator weights remains a challenge.

(2) A safety risk assessment method based on probabilistic transient stability

This method first requires fault modeling of faulty components, and then uses probability theory to solve them, represented by analytical methods, Monte Carlo methods, etc. The analytical method has high modeling accuracy, but is greatly affected by the scale of the power grid, and is prone to dimensionality disasters in large and complex power grids, resulting in the analysis method being only suitable for safety evaluation of small and simple power grids. The Monte Carlo method can overcome the shortcomings of traditional Monte Carlo methods and is suitable for safety evaluation of large and complex power systems. However, its calculation speed is linearly related to the scale of the power grid, so its solving speed will be greatly extended.

3.2 Evaluation Process of Event Driven Risk Indicators

The event driven risk assessment method consists of three parts: equipment outage rate calculation, power grid state generation, and severity calculation. This project is based on power files, XML (eXtensible Markup Language) files, etc. By extracting time-varying fault rates to obtain a set of system states, and simulating them through power files, XML files, etc., the evaluation results of fault severity are obtained.

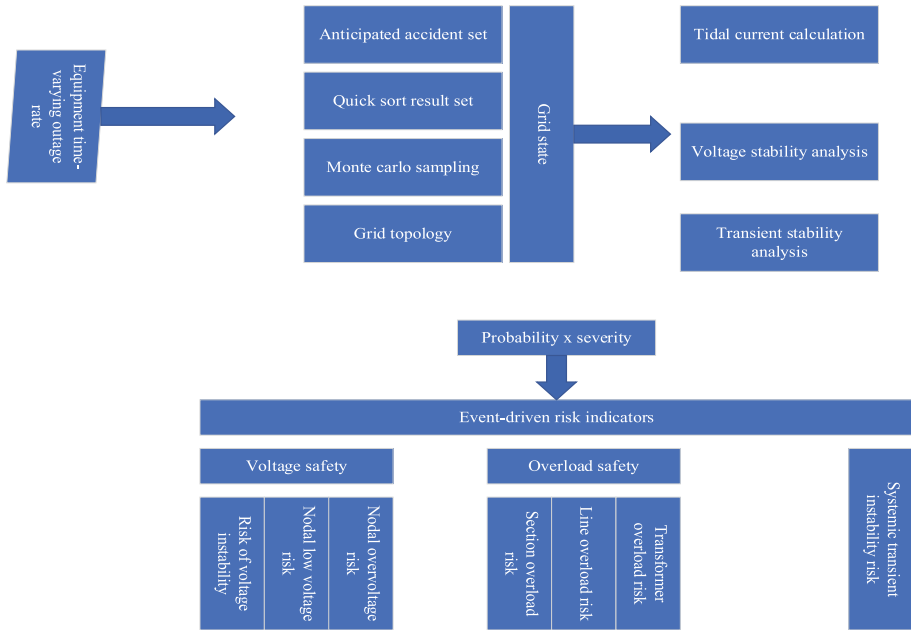


Fig. 1. Evaluation process of event driven risk indicators

Firstly, we need to construct a time-varying facility shutdown model similar to the one shown in Fig. 1. A detailed analysis of the physical mechanisms by which uncertain factors affect equipment can be conducted. For those with clear mechanisms of action, physical analysis models can be directly used. For those that are difficult to analyze physically, mathematical models such as Bayesian statistics, support vector machines, and regression prediction can be used. For physical mechanisms, stress intensity interference models can be semi quantitatively used.

3.3 Structural Learning of Bayesian Networks

This article conducts structural learning on the risk factors of power grid, equipment, and personnel accidents, and based on this, models the causes of accidents using Bayesian networks. The model can be modified multiple times by analyzing the correlation between various factors and expert opinions, and finally form a Bayesian network structure diagram, as shown in Fig. 2. From Fig. 2a, it can be seen that the Bayesian network

model for the causes of power system accidents is composed of 15 nodes, where nodes represent variables, and the connections between nodes represent direct interactions between variables. Among them, the direct influencing factors of power grid accidents (A) are inadequate inspections (P3), poor maintenance quality (T4), and equipment failures (M5). From Fig. 2b, it can be seen that the Bayesian network is an important factor, which includes 15 nodes. Among them, poor management (P2), inadequate inspection (P3), and poor construction quality (T2) are the direct factors causing equipment accidents (B). From Fig. 2c, it can be seen that the Bayesian network is an important influencing factor composed of 17 nodes. The direct influencing factors of personnel accidents (C) include insufficient safety awareness (M4), poor management (P2), poor construction quality (T2), inadequate supervision (M1), and acceptance failure (M2).

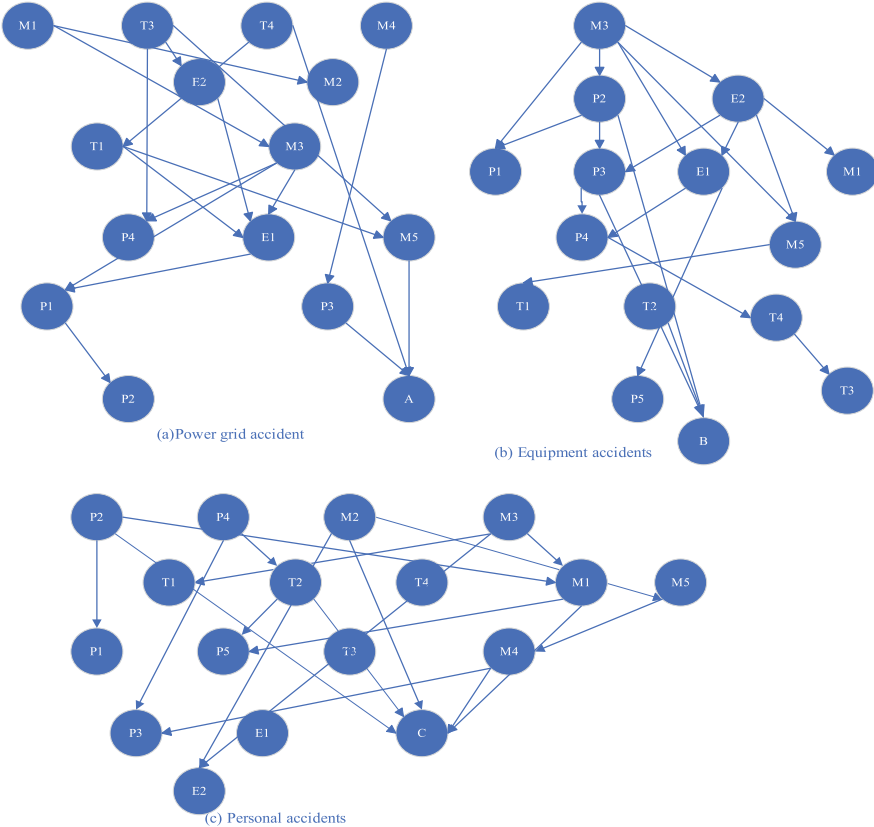


Fig. 2. Bayesian network model for accident causation

Bayesian networks are generally used for reasoning about uncertainty in knowledge systems. In cases where the problem size is small, the Bayesian network structure can be constructed by domain experts based on empirical knowledge. As the number of network nodes increases, the relationships between node variables become more complex.

Therefore, constructing Bayesian network structures from data samples has become a research focus. The structure learning method based on rating search is to select a rating function to evaluate the fitting degree of a Bayesian network structure B to the dataset U , and find the structure with the best rating as the final Bayesian network structure.

The Bayesian scoring function is based on Bayesian theorem, assuming that the prior probability of network structure B is $P(B)$. According to Bayesian theorem, the posterior probability of network structure B is as follows:

$$P(B|U) = \frac{P(B)P(U|B)}{P(U)} \quad (1)$$

Since $P(U)$ is independent of the candidate network structure, the optimal Bayesian network structure is the one that maximizes the value of $P(B)P(U|B)$. Perform the following logarithmic transformation on $P(B)P(U|B)$:

$$\log P(B)P(U|B) = \log P(B) + \log P(U|B) \quad (2)$$

In the formula: $P(B)$ refers to the prior distribution of the structure, generally assumed to be a uniform distribution, and $P(U|B)$ refers to the edge likelihood function. Therefore, maximizing the Bayesian scoring function means maximizing $P(U|B)$. If it is assumed that the parameters of the Bayesian network follow a Dirichlet prior distribution, the scoring function is obtained.

$$f_{BD}(B, U) = \sum_{i=1}^n \sum_{j=1}^q \left[\log \frac{\Gamma(a_{ij*})}{\Gamma(a_{ij*} + N_{ij*})} + \sum_{k=1}^{r_i} \log \frac{\Gamma(a_{ijk} + N_{ijk})}{\Gamma(a_{ijk})} \right] \quad (3)$$

In the formula, a_{ijk} refers to the hyperparameter value of the Dirichlet distribution,

$$a_{ij*} = \sum_{k=1}^{r_i} a_{ijk}, N_{ij*} = \sum_{k=1}^{r_i} N_{ijk}.$$

4 Results and Discussion

This article obtains representative scenarios of different wind speeds and directions through scene generation and reduction, and inputs 60 combined sample scenarios into an event driven model to obtain uncertain multi scenario power grid safety forms. The difference in the severity of some data loads is due to the prediction errors in wind speed and direction, which result in different transmission line and fiber optic fault times predicted by the equipment fault model.

From Fig. 3 above, it can be seen that using this method to evaluate indices at different standard levels yields the error between the comprehensive risk value and the true value. The results indicate that the method proposed in this article (based on Bayesian network model) has high accuracy in evaluating risks.

As shown in Fig. 4, compared with the other two methods, the Bayesian network model proposed in this paper has higher accuracy in evaluating the safety risk level of the current distribution network, and can provide a more accurate assessment. In the analysis of the sensitivity of power grid safety risks using Monte Carlo method, the

voltage limit caused by insufficient configuration capability of the connection device was not considered, resulting in low accuracy of the calculation results; The overall evaluation accuracy of the probability prediction algorithm is the lowest, mainly because this method does not perform a breakdown operation on the system’s faults, increasing the risk of the lower limit of fluctuation frequency, so it cannot accurately evaluate the safety risk of the system. The results indicate that the distribution network security risk assessment method proposed in this article can effectively evaluate the degree of harm caused by various accidents in the power grid.

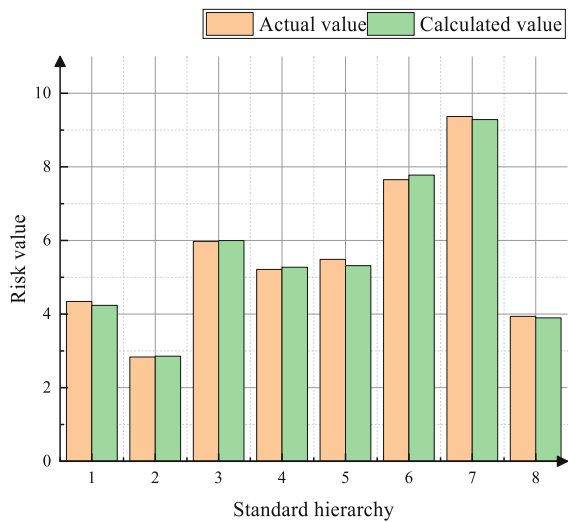


Fig. 3. Results of distribution network security risk assessment

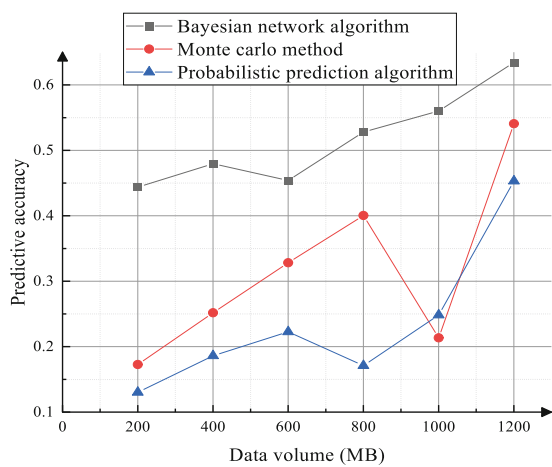


Fig. 4. Comparison of Evaluation Accuracy Results for Different Methods

From Fig. 5, it can be seen that as the training set size increases, the training time for security alarm events using different methods continues to increase. The time consumption of the Bayesian network algorithm's power grid security warning model is the smallest among the three methods, indicating that the efficiency of the model is relatively high.

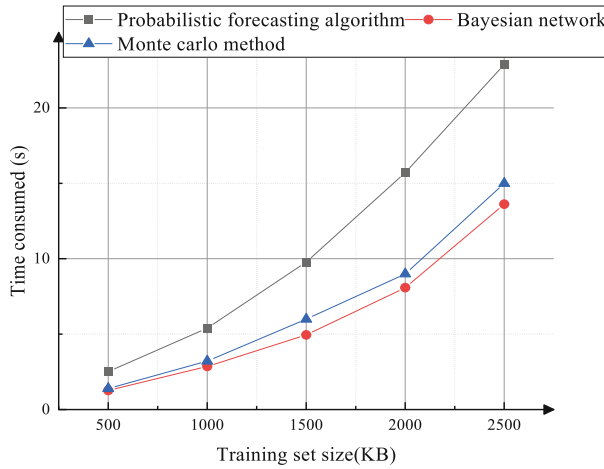


Fig. 5. Comparison of model training time for security alarm events

Table 1. False alarm rates for different methods

Algorithms	False alarm rate for low load conditions	False alarm rate for medium load condition	False alarm rate for high load condition
Bayesian networks	0.03	0.05	0.08
Monte carlo method	0.06	0.11	0.15
Probabilistic prediction	0.05	0.09	0.12

Table 1 shows the false alarm rates of three algorithms under different load conditions. It can be seen that under low load conditions, the false alarm rates of Bayesian networks and probability prediction algorithms are relatively low, while under high load conditions, the Monte Carlo method has the highest false alarm rate. Such data can more comprehensively reflect the performance differences of different algorithms in practical applications.

5 Conclusions

With the continuous development of the power system and the improvement of intelligence, power grid safety event monitoring has become an important link to ensure the stability and safety of power system operation. Power grid safety incidents may be caused by various factors, such as equipment failures, human operational errors, natural disasters, etc. These events may have serious impacts on the power grid and even lead to power outages. Therefore, timely and effective monitoring of power grid safety events is of great significance for improving the reliability and safety of power grid operation. In this study, this article delves into the monitoring technology of power grid security events and focuses on the application of Bayesian networks in this field. Through the study of risk management based security risk assessment methods, probabilistic transient stability based security risk assessment methods, and the evaluation process of event driven risk indicators, this paper reveals the important role of Bayesian networks in power grid security event monitoring.

References

1. Alavikia, Z., Shabro, M.: A comprehensive layered approach for implementing internet of things-enabled smart grid: a survey. *Digit. Commun. Netw.* **8**(3), 388–410 (2022)
2. Bi, Z., et al.: Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). *Int. J. Prod. Res.* **61**(12), 4004–4021 (2023)
3. Venu, D.N., Arun Kumar, A., Vaigandla, K.K.: Review of Internet of Things (IoT) for future generation wireless communications. *Int. J. Modern Trends Sci. Technol.* **8**(03), 01–08 (2022)
4. Akhter, R., Sofi, S.A.: Precision agriculture using IoT data analytics and machine learning. *J. King Saud Univ. Comput. Inf. Sci.* **34**(8), 5602–5618 (2022)
5. Kumar, R.L., et al.: A survey on blockchain for industrial internet of things. *Alex. Eng. J.* **61**(8), 6001–6022 (2022)
6. Sharma, S., Verma, V.K.: An integrated exploration on internet of things and wireless sensor networks. *Wirel. Pers. Commun.* **124**(3), 2735–2770 (2022)
7. Al-Khatib, A.W.: Internet of things, big data analytics and operational performance: the mediating effect of supply chain visibility. *J. Manuf. Technol. Manag.* **34**(1), 1–24 (2022)
8. Issa, W., et al.: Blockchain-based federated learning for securing internet of things: a comprehensive survey. *ACM Comput. Surv.* **55**(9), 1–43 (2023)
9. Khan, I.H., Javaid, M.: Role of Internet of Things (IoT) in adoption of Industry 4.0. *J. Ind. Integration Manag.* **7**(04), 515–533 (2022)
10. Kishor, A., Chakraborty, C.: Artificial intelligence and internet of things based healthcare 4.0 monitoring system. *Wirel. Pers. Commun.* **127**(2), 1615–1631 (2022)
11. Islam, M.M., et al.: Internet of things: device capabilities, architectures, protocols, and smart applications in healthcare domain. *IEEE Internet Things J.* **10**(4), 3611–3641 (2022)
12. Rahmani, A.M., Bayramov, S., Kiani, K.B.: Internet of things applications: opportunities and threats. *Wireless Pers. Commun.* **122**(1), 451–476 (2022)
13. Bhoi, S.K., et al.: IoT-EMS: An internet of things based environment monitoring system in volunteer computing environment. *Intell. Autom. Soft Comput* **32**(3), 1493–1507 (2022)
14. Babangida, L., et al.: Internet of things (IoT) based activity recognition strategies in smart homes: a review. *IEEE Sens. J.* **22**(9), 8327–8336 (2022)
15. Huo, R., et al.: A comprehensive survey on blockchain in industrial internet of things: motivations, research progresses, and future challenges. *IEEE Commun. Surv. Tutor.* **24**(1), 88–122 (2022)

16. Shaikh, F.K., et al.: Recent trends in internet-of-things-enabled sensor technologies for smart agriculture. *IEEE Internet Things J.* **9**(23), 23583–23598 (2022)
17. Chi, Y., et al.: Knowledge-based fault diagnosis in industrial internet of things: a survey. *IEEE Internet Things J.* **9**(15), 12886–12900 (2022)
18. Kasilingam, D., Krishna, R.: Understanding the adoption and willingness to pay for internet of things services. *Int. J. Consum. Stud.* **46**(1), 102–131 (2022)



Construction of an Online Autonomous Learning Model Based on Artificial Intelligence ChatGPT

Aimin Li and Chenming Yang^(✉)

Shandong Institute of Commerce and Technology, Jinan 250103, Shandong, China
xlyx789789@163.com

Abstract. As information technology advances swiftly, AI is increasingly deployed across numerous sectors, especially in the education industry. In order to make up for the shortcomings of traditional education, online learning has become an emerging way of learning, and there are various online learning platforms, however, the existing online learning system cannot satisfy the individualized learning requirements of learners., thus making the learning effect poor. This paper proposes an online independent learning model based on ChatGPT technology, which offers tailored learning materials to students and enhances their learning outcomes by leveraging natural language processing and AI-driven data analytics. ChatGPT's advanced data processing capability can integrate students' learning resources and adapt the learning journey in real-time based on students' behaviors to offer bespoke educational trajectories. And can be used for learning based on students' learning behaviours. ChatGPT's advanced data processing capability can integrate students' learning resources and modify the learning pathway in response to students' interaction patterns., providing students with personalized learning paths, customising learning plans according to students' needs to increase their interest in learning. The effectiveness of the new model has been verified through experiments, and the model can markedly enhance students' academic achievements and enthusiasm for learning, and the results of the satisfaction survey show that students have a high degree of satisfaction with the model, especially the personalised recommendation function and the instant feedback function of the model have been highly evaluated. The model is still in the preliminary experimental stage, and the next step will be to optimise the model algorithm to further improve the performance of the model and provide assistance for the development of educational technology.

Keywords: Artificial Intelligence · ChatGPT · Online Learning · Autonomous Learning Models · Natural Language Processing · Personalised Learning

1 Introduction

ChatGPT has been equipped with extensive language capabilities through deep learning algorithms, drawing from substantial datasets, and is capable of understanding and generating high-quality text. Applying this technology to an online learning environment

enables dynamic adjustment of course content and difficulty, providing a personalised learning experience based on the learner's ability level, interest preferences and learning progress. In addition, through natural language interaction, instantly answering students' questions and providing necessary guidance can enhance the interactivity of the learning process, as well as promote communication among learners, establish virtual communities, and enhance the social attributes of learning.

Constructing an online self-directed learning model based on ChatGPT holds substantial value for elevating the caliber of e-learning experiences, which not only helps to improve learner engagement and performance, but also helps to promote the development of educational technology, making it more humane and efficient. In addition, such research helps to narrow the digital divide and make high-quality educational resources available to more people, reflecting the great potential of technology in modern education.

2 Related Works

In the review of studies exploring the construction of online self-directed learning models based on artificial intelligence, the following literature provides valuable perspectives and research results: Zhao Yaguo [1] proposed a personalised learning model based on Bayesian knowledge tracking in his research, which dynamically adjusts the learning content and difficulty by analysing students' learning behaviour and knowledge mastery in order to achieve personalised teaching and learning, and the importance of Bayesian knowledge tracking in accurately assessing students' knowledge level. Zhong Zhuo [2] used artificial intelligence technology, machine learning and data mining to analyse and process learning data so as to learning strategies, demonstrating the potential of artificial intelligence in enhancing learning efficiency and learning experience, and providing a new perspective on the development of smart education. Yuchao Zhang [3] explored the construction of self-directed learning model in a blended learning environment, encouraging students to construct a knowledge system through self-directed exploration and practice under the guidance of teachers. Liu Yujia [4] focuses on the design of a learning model for developing students' critical thinking under the perspective of deep learning, and guides students to think deeply and analyse critically by designing challenging problems and tasks in order to cultivate their critical thinking skills, which provides a new way of thinking about instructional design in deep learning environments, and emphasises the core position of critical thinking in students' lifelong learning. Xu Chaoyi and Zhang Bo [5] focus on the cultivation of college students' online independent learning ability and its influencing factors, and through constructing an influencing factor model and conducting empirical research, they reveal the association between independent learning ability and factors such as learning motivation, learning strategy and learning environment, provide strategic suggestions for the cultivation of online independent learning ability, and emphasise the importance of the integrated effect of multiple factors in enhancing students' independent learning ability. Yang Fan [6] explored the construction of online independent learning model supported by visualisation technology, using visualisation technology to enhance students' learning experience, helping students to better understand and master knowledge by graphically

displaying learning data and learning progress, providing innovative ideas for the design of e-learning systems, emphasising the role of visualisation technology in promoting students' independent learning. Holken et al. [7] explored the interaction between bodily experience and narrative self, and developed a dynamical model of the self-schema within the framework of LIDA (Learning Intelligent Distributed Agent), which demonstrates how bodily sensations affect an individual's narrative construction and his/her self-perception by simulating human cognitive processes, providing a new perspective for understanding the formation of self-concept and its dynamics in the cognitive system provide new perspectives. Tian et al. [8] proposed a model of English independent learning based on computer network-assisted teaching and learning, using network technology, integrating multimedia resources and interactive platforms to promote students' independent learning ability, effectively increase students' learning interest and participation, and achieve remarkable results in improving English level, which provides valuable references for the innovation of teaching methods. Lai et al. [9] investigated the behavioural patterns of university students using mobile technology in self-directed language learning by integrating behavioural prediction models. Mobile tech has significantly contributed to enhancing students' learning adaptability and self-direction, but also pointed out some limitations, such as device availability and technological proficiency, which provided empirical evidence for understanding the application of mobile technology. Al-Adwan et al. [10] extended the UTAUT model to explore the role of learning traditions in understanding the intentions of continuous use of LMS, learning traditions significantly influence users' intention to continue using LMS, providing a new theoretical framework for explaining the acceptance of technology in different cultures. Xia et al. [11] explored how gender disparities and satisfaction of needs influence self-regulated learning facilitated by AI. Gender and need satisfaction had a significant impact on the effectiveness of AI-assisted learning, especially for those students who were satisfied with autonomy and relevance, which provides an important theoretical support for the development of personalised learning systems. Su et al. [12] introduced an integrated Multi-task Information Enhancement Recommendation model tailored for Self-Directed Learning Systems, enhancing educational recommendations through the synthesis of multifaceted task data, the accuracy and relevance of the recommendation system were improved. By integrating information from multiple tasks, the accuracy and usefulness of the recommendation system is improved, thus enhancing the learning experience of students, and the model is capable of significantly enhancing students' learning productivity and contentment. Li et al. [13] reconceptualized autonomous learning within the generative AI framework, examining its potential in language acquisition through an exploratory analysis, revealed how generative AI can provide language learners with innovative tools and techniques for a personalised and efficient learning experience. Kumar et al. [14] introduced a workload prediction model for cloud resource management, predicated on the principles of self-directed learning, which uses historical data to predict future workloads and thus optimise resource allocation, which can effectively reduce the cost and improve the quality of cloud services. It can effectively reduce the cost of cloud services and improve the quality of service, which provides a new solution for resource management in cloud computing. Wan and Yu [15] developed a recommendation system grounded in the Adaptive Learning Cognitive Map Model, assessing its

impact on learning outcomes, which provided personalized recommendations according to students' cognitive characteristics and learning progress, significantly improved the learning effect, and provided a new idea for the development of personalized learning recommendation system. Esiyok et al. [16] examined the acceptance of AI Chatbots for educational use among undergraduates in relation to their ICT Self-Efficacy, finding a strong correlation between the two. This implies that enhancing students' ICT proficiency is vital for fostering the adoption of AI Chatbots. Lee and Kim [17] developed a Korean language learning program for dyslexic students that combines speech recognition and handwriting recognition technologies to provide a personalised learning experience, and this AI-supported learning approach significantly improves the performance of dyslexic students in Korean language learning, especially in syllable recognition, pronunciation, and short-term memory. Lasfeto and Ulfa [18] evaluated the effectiveness of different online learning strategies by constructing a fuzzy logic model based on Fuzzy Expert Systems and Self-Directed Learning Readiness, and evaluated the effectiveness of different online learning strategies by combining fuzzy logic and self-directed learning readiness. By constructing a fuzzy logic model, the effectiveness of different online learning strategies was evaluated, and the strategy combining fuzzy logic and Self-Directed Learning Readiness (SDLR) can significantly improve students' learning performance, which provides a theoretical basis for the design of online education. Li [19] developed an online distance education system using intelligent agents to provide personalised learning suggestions and services to support distance learners' independent learning, effectively improving distance learners' engagement and learning efficiency, providing new ideas for the innovation of distance education technology. Kadhim et al. [20] proposed a new self-directed learning framework for clustering integration, which combines the concepts and techniques and aims to improve the performance of clustering algorithms, especially when dealing with large-scale datasets showing better clustering results, providing a new methodological contribution to the field of data mining and machine learning. Jun and Min [21] focused on the tuning of generative AI parameters in online self-directed learning environments and explored the impact of different parameter settings on the response quality of generative AI. Reasonable parameter configurations can significantly improve the performance of generative AI in terms of code generation and annotation creation as well as provide more emotionally supportive feedback, which provides a practical guide to optimise generative AI for use in educational technology.

Upon examining the literature, it is clear that current research has carried out extensive work and in-depth discussions on self-directed learning, AI-assisted education, and learning management systems, which not only provide rich theoretical and empirical support for the development of educational technology, but also provide new ideas and methods for future educational practices. These studies provide a multi-dimensional perspective and rich empirical data for the construction of online self-directed learning models based on artificial intelligence, and provide valuable references for subsequent research and practice.

3 Methods

3.1 System Design

The core of the design of the ChatGPT-based online self-directed learning model lies in the integration of advanced natural language processing technologies and pedagogical theories in order to provide a more personalised and efficient online learning experience. The system architecture consists of three key components: a user interface module, an intelligent recommendation engine, and a data analysis platform, as shown in Fig. 1.

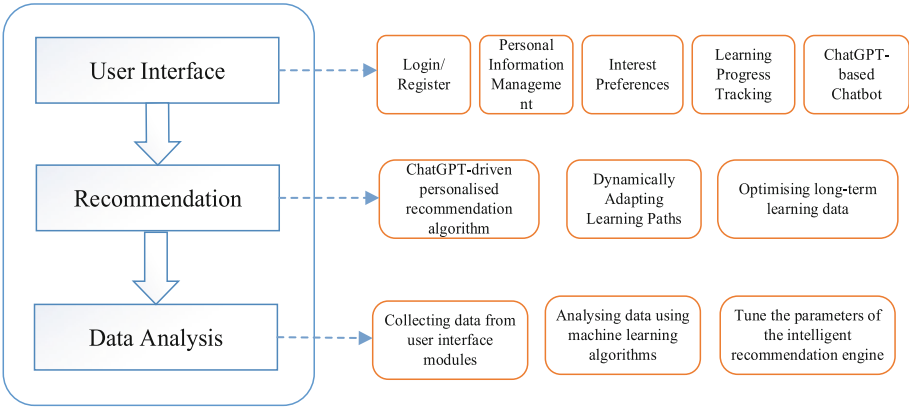


Fig. 1. System Architecture Diagram

The main function of the Interaction Design component is to communicate with learners. The module obtains information about the user’s profile, interests and learning progress through a well-designed user interface, provides a friendly and easy-to-operate system interface for the user, enables the learner to track and manage his/her own learning status with ease, and answers the user’s queries through an interactive chat tool with embedded ChatGPT technology. The system also provides instant feedback and tutorials based on student learning to enhance user engagement and learning experience.

The Intelligent Recommendation Engine is the core of the whole system, using ChatGPT’s powerful text processing and natural language parsing capabilities to tailor-make course recommendations for learners, modifying educational pathways on-the-fly based on learners’ performance to guarantee incremental progress at their individual pace and skill level, and at the same time optimising the recommendation algorithms based on cumulative learning data in order to promote the continuous growth of learners.

The data analysis system is responsible for collecting data from the students’ learning process and conducting in-depth processing and analysis, using machine learning and big data technologies to identify learners’ behavioural trends and optimise the performance of the intelligent recommendation system accordingly to improve the relevance and effectiveness of the recommendations, tracking the learning activities of the learners in real time, accurately grasping the learners’ needs and adjusting the recommendation strategy accordingly to enhance the learning effectiveness.

3.2 Pre-trained Convolutional Neural Networks

In order to test the effectiveness of the online self-directed learning model based on ChatGPT, a series of experiments are designed in this paper, and the experimental flow is detailed in Fig. 2.

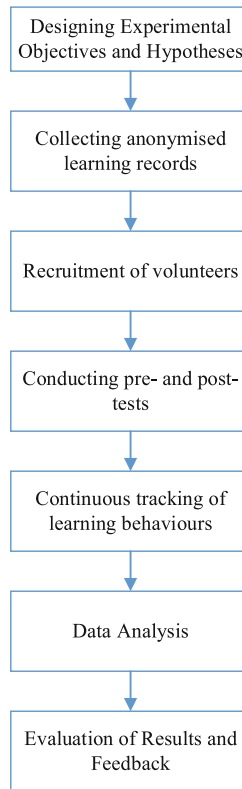


Fig. 2. Experiment flow chart.

After defining the problem to be addressed by the study, the next step is to prepare the resources required for the experiment, which includes obtaining the necessary data from reliable online learning platforms, which are anonymised to protect user privacy, as well as ensuring that the information collected reflects the participants' learning behaviours and outcomes. The researchers will then screen eligible participants based on certain criteria and randomly assign them to the experimental and control groups to ensure that the two groups are comparable before the start of the experiment.

A pre-test is conducted before the experiment officially starts to record the basic situation and initial level of the participants, which is crucial for the subsequent evaluation of the effects. During the experiment, participants in the experimental group will be guided to use the ChatGPT-based self-directed learning tool to enhance their learning efficiency, while the control group will continue to use the traditional learning method.

Throughout the experiment, not only the changes in the final scores will be followed, but also the participants’ learning habits, level of engagement, and completion of tasks will be continuously tracked, and these dynamic data will help to understand the actual effects of the experimental tools more comprehensively.

Upon experiment conclusion, statistical analyses will assess the significance of differences between the experimental and control groups via a pre-post comparison, while qualitative analyses, such as questionnaires or in-depth interviews, will be used to obtain the participants’ perceptions and feelings about the new learning tool.

4 Results and Discussion

4.1 Experimental Design

A total of 200 students aged between 15 and 25 years old from different disciplines in high school and university participated in this experiment to ensure a broad and diverse sample. The participants were randomly assigned into two groups: experimental and control, 100 in each group, and their basic information is detailed in Table 1.

Table 1. Participant Demographics

Group	Sex (m/f)	Age (years)	Education Level	Subject Areas
Experimental Group	55/45	15–25	High School/University	Maths/Physics/Chemistry/Biology
Control group	50/50	15–25	High School/University	Maths/Physics/Chemistry/Biology

Prior to the commencement of the experiment, all participants were subjected to the same baseline test to assess their initial level of knowledge in the chosen subject area and the results are shown in Table 2.

Table 2. Distribution of baseline test scores

Group	Number of participants	Mean score	Highest score	Lowest score	Standard deviation
Experimental Group	100	65	85	50	5.2
Control group	100	64	84	49	5.1

Learners in the experimental cohort were assigned to use the ChatGPT-based online self-directed learning platform for one month, during which they could access a variety of learning resources provided by the platform and interact with the platform’s in-built

chatbot for instant feedback. Students in the control group continued to use a traditional online learning platform that did not include personalised recommendations or live support features. The post-experiment test scores are shown in Table 3.

Table 3. Distribution of post-test scores after the experiment

Group	Number of participants	Mean score	Highest score	Lowest score	Standard deviation
Experimental Group	100	78	95	60	6.3
Control group	100	70	88	55	5.5

One month later, all participants were given the same test again to assess their learning outcomes. In addition, we collected data on the participants' learning behaviours, including the frequency of logging in, the time spent studying, and the number of tasks completed, the system automatically records the students' learning behaviours, as shown in Table 4.

Table 4. Statistical data on learning behaviour

Group	Frequency of logging in (times/week)	Average study time (hours/day)	Number of tasks completed (units/week)
Experimental Group	5.5	1.75	12
Control group	4.2	1.4	8.5

In order to gain insight into students' perceptions of the new model, a questionnaire was administered to all participants, asking them about their satisfaction with the learning experience, problems encountered, and suggestions for improvement. Students were surveyed on their satisfaction with their learning experience, and Table 5 demonstrates the results of the survey feedback on students' satisfaction with their learning.

Table 5. Results of the Survey on Learning Satisfaction

Group	Very satisfied (%)	Satisfied (%)	Neutral (%)	Dissatisfied (%)	Very dissatisfied (%)
Experimental Group	45	35	15	3	2
Control group	25	40	25	7	3

4.2 Results

By analysing the results of the experimental and control groups, it can be seen that there was a very significant improvement in the performance of the students in the experimental group, with the average score increasing from 65% in the initial test to 78% after the experiment, while in comparison, the score of the students in the control group only increased from 64 to 70%, such a result shows that the experimental group's learning scores have increased by about 20% points, while the control group's increase is about 9. The results show that the experimental group's academic scores increased by about 20% points, compared to about 9% points for the control group. In addition to the improvement of learning scores, the learning behaviours of the students in the experimental group also showed significant positive changes, with the frequency of logging in the experimental group increasing by about 30% points, the average length of study per day increasing by as much as 25% points, and the amount of completed tasks being nearly 40% points more than that of the control group, which indicates that the online self-directed learning model based on ChatGPT not only improves learning scores, but also increases students' motivation and improves the learning outcomes of the students. Students' motivation and increased their interest in learning. In the survey document feedback experiment, most of the students in the experimental group reflected that the customised study plans and instant feedback provided by the platform were great, which not only helped them to understand the difficulties and key points of learning in a deeper way, but also increased their motivation to learn, while comparatively, the students in the control group pointed out the inadequacy of the personalised support more often.

Through these specific experimental data, it can be concluded that the online self-directed learning model based on ChatGPT shows obvious advantages in improving learning performance, increasing learning interest, and promoting students' active learning, etc., and has been well evaluated and welcomed by the students, which provides a solid basis for further improvement and refinement of this type of learning model in the future.

4.3 Discussion

The ChatGPT-supported self-directed learning model significantly improved student performance (20%) and engagement in the experimental group, outperforming the control group (9%). Student feedback showed that the personalisation and instant feedback features of the model were well received. The new model performed better in terms of personalisation and instant interaction compared to traditional online learning. Limitations of the study include the short-term experimental period and insufficient sample diversity, and further research is needed to improve the depth and breadth of learning content.

Compared with traditional learning paths, the online self-directed learning model based on ChatGPT shows stronger adaptability and flexibility. Traditional learning paths are often pre-set course sequences, and students learn according to a set schedule and content order, which is a one-size-fits-all approach that ignores the different starting points, learning speeds, and interest preferences of each learner, leading to the fact that some students encounter difficulties in certain knowledge points but are unable to get

timely help, thus affecting the overall learning effect. This one-size-fits-all approach ignores the different starting points and interest preferences of each learner, resulting in some students encountering difficulties in certain knowledge points but not being able to get help in time, which in turn affects the overall learning effect. When the system detects that a student is not performing well in a particular area, it will automatically adjust the subsequent learning plan by adding more relevant practice questions or reading materials to consolidate the student's understanding. In addition, by analysing students' learning behaviour patterns, the ChatGPT model also identifies which topics are students' strengths and weaknesses, and adjusts the course difficulty accordingly to ensure that each student learns at his or her optimal level of difficulty. This customised learning path not only improves the efficiency of learning, but also makes it fun for students and enhances their motivation to learn.

Another significant advantage is the feature of instant feedback and support. The chatbot built into the ChatGPT model is able to answer students' questions and provide instant feedback in real time, just like a real teacher. This kind of instant interaction is unmatched by traditional eLearning platforms. Traditional platforms usually rely on pre-recorded videos or static text materials, students can only passively receive information, once they do not understand something, they often need to wait for the teacher's reply or consult other materials to solve the problem, while in the ChatGPT model, students can ask questions at any time and get answers immediately, which not only eliminates the confusion of the students in a timely manner, but also stimulates their curiosity to continue to explore the unknown curiosity. Instant interaction is not only limited to answering questions and solving puzzles, but also includes instant correction and feedback on students' homework. Through dialogue-based communication with students, the ChatGPT model can better understand students' ideas, provide more targeted advice, and help students gradually build up their problem-solving skills, which not only improves learning efficiency, but also makes students feel valued and supported, and enhances their confidence in learning.

In addition to focusing on the final learning results, the ChatGPT model also focuses on the all-round monitoring of the students' learning process. By continuously tracking dynamic data such as students' log-in frequency, daily study time and task completion, the model can help teachers or learning administrators to have a comprehensive understanding of the students' learning status and intervene when necessary. If the system finds that a student has not logged in to the learning platform for several consecutive days, it may send an early warning to the teacher, reminding him or her to follow up in time to find out if the student has encountered any difficulties. By analysing this data, educators can more accurately determine which teaching methods are effective and which areas need improvement, so as to continuously optimise teaching strategies. In contrast, traditional online learning platforms often focus only on the final test scores, while ignoring some key indicators in the learning process. Learning is a dynamic development process, relying only on the results of a test to judge the effectiveness of student learning is not comprehensive, by carefully observing the learning behaviour of the students, the ChatGPT model provides teachers with more dimensions of data support to help them make more scientific teaching decisions.

The ChatGPT-based online self-directed learning model shows obvious advantages in personalized learning path design, instant feedback support and learning behavior monitoring, etc. These features not only improve the learning effect, but also greatly enrich the online learning experience. Through continuous research and optimization, this model is expected to bring more efficient and personalized learning experience to more learners in the future.

5 Conclusion

Although ChatGPT-driven online self-directed learning models show potential in real-world applications, challenges remain. Iteration of technology requires continuous updating of the model to keep up with the latest advances, the long-term effects of the model need to be further validated due to the short duration of the study, the current study focuses on adolescents and future research should be expanded to a wider range of age groups and cultural backgrounds, ensuring the quality and depth of the learning content is key, and technological development should not affect the essence of education.

To address these challenges, firstly, it is recommended that a long-term research tracking mechanism be established to continuously assess the long-term impact of modelling on student learning outcomes. Second, due to the short experimental period, further research is needed to verify the effectiveness of the model over a longer time span. In addition, the current study focuses on the adolescent student population and needs to be expanded to cover a wider range of age groups and cultural backgrounds in the future. Finally, it is equally important to ensure the quality and depth of the learning content, and it must be ensured that technological advances do not sacrifice the core values of education.

To address these challenges, the following measures can be taken: first, establish a long-term tracking research mechanism to continuously monitor the impact of the model on students' long-term learning outcomes; second, conduct diversified user studies to ensure that the model is applicable to different types of user groups; third, strengthen the content review and updating mechanism to ensure that the learning materials are accurate and up-to-date; and fourth, strengthen the technical support team to respond to technical updates and user feedback to ensure the stability of the model and user experience.

In conclusion, the online self-directed learning model based on ChatGPT shows great potential in enhancing the learning effect and improving the learning experience. Through continuous research and improvement, this model is expected to bring more efficient and personalised learning experience to more learners, and promote the development of online education in a more intelligent direction.

References

1. Yage, Z.: Construction and Application of Personalised Learning Model Based on Bayesian Knowledge Tracking. Henan Normal University (2023). <https://doi.org/10.27118/d.cnki.ghesu.2023.000339>

2. Zhuo, Z.: Research on the Construction and Application of Intelligent Learning Model Supported by Artificial Intelligence. Northeast Normal University (2023). <https://doi.org/10.27011/d.cnki.gdbsu.2023.000004>
3. Zhang, Y.: Research on the Construction of Self-directed Learning Model in Blended Learning Environment. Jilin University (2022). <https://doi.org/10.27162/d.cnki.gjlin.2022.003073>
4. Yujia, L.: Research on the Design of Learning Model for Developing Students' Critical Thinking Under the Perspective of Deep Learning. Central China Normal University (2022). <https://doi.org/10.27159/d.cnki.ghzsu.2022.001468>
5. Chaoyi, X., Bo, Z.: Model and empirical research on the influencing factors of college students' online independent learning ability cultivation. *J. Xichang Coll. (Nat. Sci. Ed.)* **35**(04), 98–101 (2021). <https://doi.org/10.16104/j.issn.1673-1891.2021.04.018>
6. Fan, Y.: Construction and Empirical Research on Online Independent Learning Model Supported by Visualisation Technology. Northeast Normal University (2021). <https://doi.org/10.27011/d.cnki.gdbsu.2021.001731>
7. Holken, A., Kugele, S., Newen, A., Franklin, S.: Exploring the interplay of embodied and narrative selves: self-pattern dynamics in the LIDA framework. *Cogn. Syst. Res.* **81**, 25–36 (2023). <https://doi.org/10.1016/j.cogsys.2023.03.002>
8. Tian, M., Fu, R., Tang, Q.: Development of a network-assisted English learning model focusing on autonomy. *Comput. Intell. Neurosci.* **2022**, 1 (2022). <https://doi.org/10.1155/2022/8646463>
9. Lai, Y., Saab, N., Admiraal, W.: The role of mobile technology in self-directed language learning: insights from an integrative behavioral model. *Comput. Educ.* **179**, 104413 (2022). <https://doi.org/10.1016/j.compedu.2021.104413>
10. Al-Adwan, A.S., Yaseen, H., Alsoud, A., Abousweilem, F., Al-Rahmi, W.M.: Understanding prolonged engagement with learning management systems: A UTAUT model extension. *Educ. Inf. Technol.* **27**(3), 3567–3593 (2022). <https://doi.org/10.1007/s10639-021-10758-y>
11. Xia, Q., Chiu, T.K.F., Chai, C.S.: Gender and need fulfillment in AI-mediated self-regulated learning: a moderation analysis. *Educ. Inf. Technol.* **28**(7), 8691–8713 (2023). <https://doi.org/10.1007/s10639-022-11547-x>
12. Su, Y., et al.: Enhancing educational recommendations through multi-task information processing for self-directed learners. *Expert Syst. Appl.* **252**, 124073 (2024). <https://doi.org/10.1016/j.eswa.2024.124073>
13. Li, B., Bonk, C.J., Wang, C., Kou, X.: Reenvisioning self-directed learning in the age of generative AI: a language learning inquiry. *IEEE Trans. Learn. Technol.* **17**, 1515–1529 (2024). <https://doi.org/10.1109/tlt.2024.3386098>
14. Kumar, J., Singh, A.K., Buyya, R.: A self-learning approach to workload forecasting for cloud resource management. *Inf. Sci.* **543**, 345–366 (2021). <https://doi.org/10.1016/j.ins.2020.07.012>
15. Wan, H., Yu, S.: Leveraging adaptive cognitive mapping for personalized learning recommendations. *Interact. Learn. Environ.* **31**(3), 1821–1839 (2023). <https://doi.org/10.1080/10494820.2020.1858115>
16. Esiyok, E., Gokcearslan, S., Kucukergin, K.G.: Student acceptance of AI-powered chatbots in technology-enhanced self-directed learning. *Int. J. Hum. Comput. Interaction* (2024). <https://doi.org/10.1080/10447318.2024.2303557>
17. Lee, A.-J., Kim, K.-W.: Crafting an AI-driven hangeul learning solution for dyslexic learners. *J. Digit. Cont. Soc.* **23**(5), 781–791 (2022). <https://doi.org/10.9728/dcs.2022.23.5.781>
18. Lasfeto, D.B., Ulfa, S.: Fuzzy expert systems in online learning strategy modeling: impact on self-directed learning outcomes. *J. Educ. Comput. Res.* **60**(8), 2081–2104 (2023). <https://doi.org/10.1177/07356331221094249>
19. Li, R.: Advancing web-based distance education with AI-agent technology. *J. Intell. Fuzzy Syst.* **40**(2), 3289–3299 (2021). <https://doi.org/10.3233/jifs-189369>

20. Kadhim, M.R., Zhou, G., Tian, W.: Cluster ensemble framework for autonomous learning. J. King Saud Univ. Comput. Inf. Sci. **34**(10), 7841–7855 (2022). <https://doi.org/10.1016/j.jksuci.2022.07.003>
21. Jun, J.-Y., Min, Y.-A.: Fine-tuning generative ai for online self-directed learning. J. Korea Soc. Comput. Inf. **29**(4), 31–38 (2024). <https://doi.org/10.9708/jksoci.2024.29.04.031>



Application of PMS Graphic Intelligent Recognition and Analysis Based on Contact Diagram Automatic Generation Technology

Wei Ma, Qiang Li^(✉), Yuan Yao, and Xinkai Chen

Power Supply Service Command, State Grid Xinjiang Electric Power Co., Ltd., Changji Power Supply Company, Changji, Xinjiang, China
19109947778@163.com

Abstract. With the continuous expansion and increasing complexity of the power grid, the traditional method of manually drawing electrical connection diagrams is no longer able to meet the needs of real-time updates and maintenance. This article studies the application of PMS graphic intelligent recognition and analysis based on contact graph automatic generation technology, aiming to improve the efficiency of power grid construction and management. The study adopts the Generalized Chromosomes Genetic Algorithm (GCGA) algorithm, combined with the advantages of graph computing and quantum genetic algorithm, to optimize search efficiency and solution quality through adaptive crossover and mutation rates, as well as grouping competition and optimal selection mechanisms. The research includes steps such as topology data analysis, intelligent layout, and graphic rendering to ensure that the automatically generated contact diagram is both accurate and easy to understand. The experimental results show that the application of automatic generation technology for contact diagrams can achieve a maximum topology integrity of 99.8%, and the time required for data organization, layout optimization, and graphic rendering is significantly faster than traditional manual drawing methods. The system can quickly respond and accurately recover in the face of power grid changes and abnormal situations. These results demonstrate the effectiveness and practicality of automatic generation technology in improving the automation level of power grid management, reducing human errors, and supporting real-time monitoring and rapid fault response of the power grid.

Keywords: Lean Management System For Equipment Operation And Maintenance · Automatic Generation Of Contact Diagram · Intelligent Graphic Recognition · Graph Computation · Genetic Algorithm · Topology Analysis · Power System Operation And Maintenance

1 Introduction

With the rapid development of the power system, the scale and complexity of the power grid continue to increase, and traditional manual drawing of electrical connection diagrams can no longer meet the fast, accurate, and efficient needs of modern power grids.

At present, the construction and management of the power grid are facing difficulties in updating contact maps, poor information accuracy, and insufficient real-time performance. These problems not only affect the operation and maintenance efficiency of the power grid, but also increase the difficulty and risk of power grid fault handling. Therefore, researching and developing automated electrical contact diagram generation technology is of great significance for improving the operation and management level of the power grid. This article addresses the limitations of existing methods and improves the efficiency of power grid construction and management through the application of PMS (Equipment (Asset) Lean Management System) graphic intelligent recognition analysis based on contact diagram automatic generation technology. This article deeply analyzes the topology structure and equipment asset information of the power grid, studies the method of generating automated contact diagrams, and enables it to automatically and accurately generate electrical contact diagrams, providing technical support for real-time monitoring and rapid response to faults in the power grid.

Based on the GCGA algorithm for automated contact graph generation technology, combining the advantages of graph computing and quantum genetic algorithm, and through adaptive crossover and mutation rates, as well as grouping competition and optimal selection mechanisms, this article optimizes search efficiency and solution quality, and implements key technologies such as distribution network topology data analysis, model generation, intelligent layout, and graphic rendering to ensure that the automatically generated contact graph is both accurate and easy to understand. Finally, the article validates the effectiveness of the automatic generation technology based on contact diagrams through actual power grid cases. The article first introduces the research background and current situation, and elaborates on the research methods and implementation steps of automated contact diagram generation technology, and finally, demonstrates the application effect of the technology through experiments and proposes relevant improvement measures.

2 Related Work

Improving efficiency and sustainability has become a hot research topic in modern industrial production and power system management. Quiroz Flores J C explored the lean production management model of implementing preventive maintenance methods in the plastic industry through case studies to improve production efficiency [1]. Amrani M A et al. implemented an integrated maintenance management system for the biscuit industry to monitor production lines and improve production quality and maintenance efficiency [2]. Singh M implemented the Environmental Lean Six Sigma framework at a medical equipment manufacturing unit in India to improve production efficiency and reduce environmental impact [3]. Díaz Reza J R et al. studied the relationship between lean manufacturing tools and their sustainable economic benefits, and found that lean tools have a significant effect on improving economic benefits [4]. Ufa R A et al. reviewed the impact of distributed generation on the power system and pointed out the potential of distributed generation in improving energy efficiency and reducing environmental impact [5]. In the field of power systems, Sen P et al. proposed a method for automatically generating single line diagrams of distribution networks and supporting incremental

updates to improve the automation level of power grid management [6]. Asoh D A designed and implemented an automatic overvoltage and undervoltage protection system for single-phase low-voltage power lines to improve the reliability and safety of the power grid [7]. Jabbar F I et al. used artificial bee colony algorithm to optimize the detection of single-phase grounding faults in distribution networks, in order to improve the accuracy and efficiency of fault detection [8].

Although the above research has made some progress in improving production efficiency, power grid management, and fault detection, there are still some shortcomings. Existing research mostly focuses on case analysis of a single industry, lacking comprehensive comparison and application across industries. In addition, research on automation of power grid management, especially in intelligent recognition and automatic generation of contact diagrams, is not yet in-depth enough. This article aims to improve the efficiency of power grid construction and management by studying the application of PMS graphic intelligent recognition analysis based on the automatic generation technology of contact diagrams. By combining the advantages of graph computing and quantum genetic algorithm, automatic and intelligent generation of power grid contact diagrams can be achieved. Through this method, not only can the automation level of power grid management be improved, but technical support can also be provided for real-time monitoring and rapid response to faults in the power grid, thereby contributing to the efficient and sustainable development of the power system.

3 Method

3.1 Data Collection and Preprocessing

The data in this article mainly comes from the PMS system, which integrates a large amount of lean management of power equipment and asset information, providing detailed information on various components in the distribution network, including the location, status, specifications, and connection relationships of equipment such as transformers, circuit breakers, cables, and transmission lines [9]. The types of data collection include equipment ledger information, including equipment models, specifications, installation dates, and maintenance records. At the same time, topology connection data provides detailed records of the connection methods between devices, including parallel, series, or ring networks. In addition, the geographical location of the device, including latitude and longitude coordinates, is crucial for subsequent graphic rendering and spatial analysis. The real-time load, current, and voltage of the equipment are also important parts of data collection.

The collected data is first carefully cleaned to correct incorrect device numbers and mismatched link relationships; and through data conversion, the connection relationships described in the text are transformed into a structured data format for analysis and processing [10]. Finally, continuing with data normalization to ensure that all data follows agreed standards and formats. This article converts all voltage values into a unified kV unit. In terms of data consistency check, the in and out degrees of each device connection will be checked to verify whether the links are consistent.

3.2 Topology Data Analysis

This article uses graph theory methods to analyze and understand the topology structure of the power grid. Graph theory is used to simulate and analyze a network composed of nodes and edges of electrical connections between devices in a power system, including transformers, circuit breakers, etc. Figure 1 shows the topology of the distribution network.

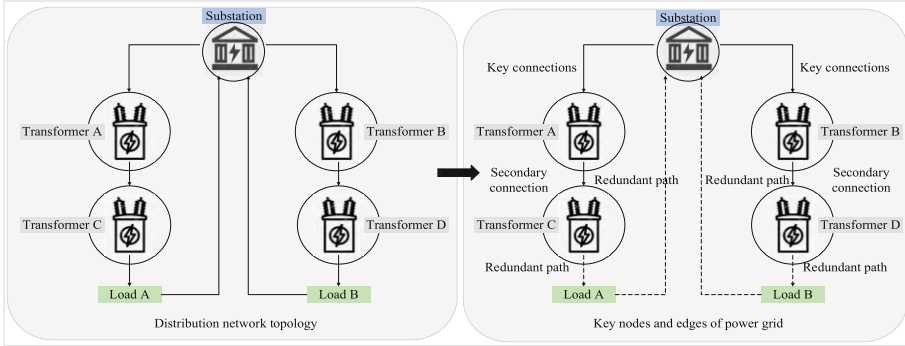


Fig. 1. Topology structure of distribution network

The raw data extracted from the PMS system in this article is used to construct a graph representation, where each power device is a node in the graph, and the connection relationships between devices are represented by edges. Graph theory is applied to analyze the topological characteristics of the network, such as connectivity, existence of loops, and degree of nodes. Depth first search and breadth first search algorithms are used to traverse the power grid map, in order to identify and verify the connection relationships between lines. By determining whether each device in the power grid can be reached by other devices, the connectivity of the power grid is verified, and the loops in the power grid are identified. In the process of topological data analysis, the basic characteristics of a graph include the clustering coefficient of nodes, which measures the tightness of connections between neighbors of a node [11]. The calculation formula is as follows:

$$C_u = \frac{2E}{k_u(k_u - 1)} \quad (1)$$

Among them, C_u is the clustering coefficient of node u , and E is the actual number of edges that exist between the neighbors of node u . k_u is the degree of node u , and a high clustering coefficient means that the neighbors of a node tend to connect to each other. In order to gain a deeper understanding of the dynamic characteristics of the power grid, this article constructed a dynamic characteristic diagram of the power grid, as shown in Fig. 2:

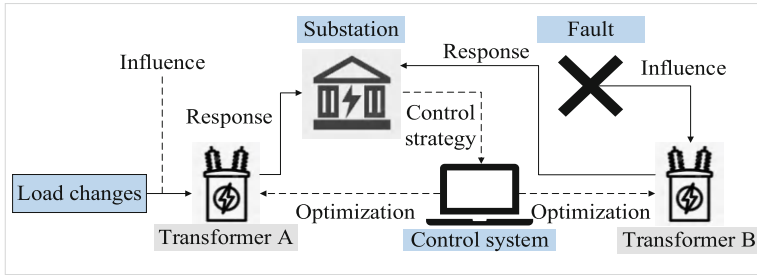


Fig. 2. Dynamic characteristics of the power grid

The dynamic characteristic diagram of the power grid provides an intuitive perspective for analyzing and understanding the behavior of the power grid in the face of different operating conditions and external disturbances. The dynamic characteristic diagram of the power grid helps identify vulnerable links in specific situations by displaying the interactions and dependencies between various components in the grid [12].

3.3 Intelligent Layout Algorithm

In the automatic mapping process of the distribution network, this article optimizes the layout of lines and nodes based on the GCGA algorithm [13] to reduce the intersection of edges and the overlap of nodes, and ensure the clarity and readability of the graphics. The algorithm combines quantum genetic algorithm and layout optimization techniques in graph theory, iteratively improving layout schemes by simulating natural selection and genetic mechanisms. Each layout scheme is considered as an individual, and the solution space of the entire layout process is composed of all possible layout schemes.

In group competition and optimal selection, the algorithm will start from a randomly generated set of initial layout schemes, each scheme including the position coordinates of all nodes. Considering the compactness of the layout, the number of intersecting edges, and the degree of node overlap, the selection operation will choose an excellent layout scheme for reproduction, and individuals with high fitness have a higher probability of being selected. Applying cross operations to combine two layout schemes generates a new layout, while mutation operations randomly adjust the positions of certain nodes to increase the diversity of the solution space.

3.4 Graphic Rendering and Generation

In the automatic mapping process of power distribution networks, graphic rendering and generation are the key final steps. During the graphic rendering process, first initializing a canvas based on optimized layout data, and then draw each power device and their connections on the canvas according to the coordinate information of nodes and edges. In order to improve the readability and aesthetics of graphics, style the nodes and edges, including selecting appropriate colors, line types, and label fonts, to ensure that the graphics are clear and have a professional appearance [14]. To improve the representation ability of the graph, interactive labeling and prompts were also applied to each node and

edge, enabling them to display more details when moving, including device types, voltage levels, load capacity, etc. After completing the drawing of nodes and edges, the entire image was globally adjusted and optimized, including scaling ratio, alignment method, and edge orientation, to meet the requirements of different display devices and output formats. Finally, outputting the obtained contact diagram as various types of graphics, including SVG, PNG, PDF, etc., for easy application on various platforms.

4 Results and Discussion

4.1 Experimental Design

This article will evaluate the performance of automatic generation technology in terms of accuracy, efficiency, and reliability through experiments, and compare it with traditional manual drawing methods. The experimental server will use the latest high-performance processors and have sufficient memory to ensure the efficient operation of the algorithm. The experimental dataset adopts distribution network models of various scales and complexities to comprehensively evaluate their performance under various operating states. These models cover from simple single line networks to complex multi ring networks, ensuring the representativeness and universality of experimental results. By comparing with the actual power grid, the effectiveness of the established models is verified, and the time of each connection diagram generation will be recorded to evaluate the efficiency of the algorithm. Finally, by simulating changes in the power grid topology, the real-time update capability and stability of the automatic generation technology are tested.

4.2 Accuracy Evaluation

This article evaluates the accuracy of the model in accurately reflecting the location, type, status, and interrelationships of devices, with node consistency, connection accuracy, and topology integrity as the main evaluation indicators. Among them, the node consistency index is used to measure whether each node in the network matches the nodes in the actual power grid. The connection accuracy index is used to evaluate whether each connection line in the automatically generated graph truly reflects the physical connection between devices. The topology integrity index is used to check whether the automatically generated connection diagram fully displays the topology of the power grid, including all necessary nodes and connections. Figure 3 shows the accuracy data of the model in different evaluation metrics.

In the accuracy data of Fig. 3, the highest node consistency reaches 98.9%, and although the accuracy of the connections is relatively average, the highest connection accuracy can also reach 97.8%. It can be clearly observed that the value of topological integrity is the highest, with the highest topological integrity being 99.8% and the lowest being 96.2%. It can be seen that the automatically generated connection diagram can effectively display the topology of the power grid, including all necessary nodes and connections, even as the scale of the power grid increases.

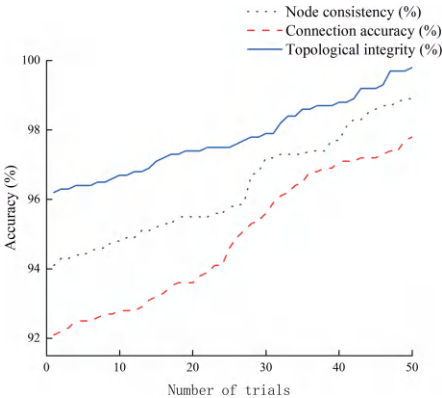


Fig. 3. Accuracy assessment

4.3 Efficiency Evaluation

In the efficiency evaluation section, this article will quantify and compare the efficiency differences between automatic generation technology and traditional manual drawing methods in drawing distribution network connection diagrams. Efficiency evaluation focuses on measuring the time required to complete the connection diagram. In each experiment, automatic generation technology and traditional methods were used to create grid connection diagrams of the same scale and complexity, and small, medium, and large-scale grid models were selected to ensure the universality of the evaluation results. Table 1 shows the specific time data collected in this article for data loading, processing, layout optimization, and graphic rendering:

Table 1. Efficiency evaluation

Grid size	Drawing process	Contact map automation (minute)	Manual drawing (minutes)
Small	Data collation	2.4	14.6
	Layout	1.5	32.8
	Drawing	1.2	17.9
Medium	Data collation	5.6	31.4
	Layout	3.8	68.1
	Drawing	2.3	34.8
Large	Data collation	10.7	62.6
	Layout	5.7	135.4
	Drawing	3.7	68.5

In Table 1, during the data organization stage, the time required for automatic generation technology is significantly less than that for manual drawing methods. For small

power grids, automation technology only takes 2.4 min, while manual methods take 14.6 min. This difference is more pronounced in medium and large power grids, where automation technology for medium power grids takes 5.6 min and manual methods take 31.4 min; large scale power grid automation technology takes 10.7 min, while manual methods can take up to 62.6 min. Automation technology has significant speed advantages in processing data and preparing drawings, and this advantage becomes more prominent as the scale of the power grid increases. In the layout and drawing stages, automation technology has also demonstrated high efficiency. In large-scale power grids, the drawing time for automation technology is 3.7 min, while manual methods take 68.5 min, mainly due to the ability of automation technology to handle repetitive tasks and accurately control graphical output.

4.4 Reliability and Stability Testing

In order to ensure that the automatically generated contact diagram technology can continuously and stably provide accurate graphical output in the face of changes and abnormal situations in the actual operation of the power grid, this article designs a simulation test of power grid changes and abnormal situations to evaluate the system's response capability and stability. In simulating common changes in the power grid, the ability of automatic generation technology to handle these changes is tested by automatically injecting new or removed power equipment, changing equipment connection relationships, adjusting line configurations, and other changes into the power grid model through scripts. For the simulation of abnormal situations, equipment failures, data loss or damage, network attacks, etc. were included, and 6 experiments were conducted for each change. Figure 4 shows the reliability and stability data under power grid changes.

In the experiment of increasing the transformer, the response time ranges from 3.2 to 3.8 s, the recovery time ranges from 5.1 to 5.8 s, and the total time ranges from 8.7 to 9.4 s. It can be observed that when the transformer is reduced by one, the overall response and recovery time also decrease accordingly. In the data of Fig. 4, the response and recovery time of the system in line configuration adjustment is relatively long, but the longest response time can also be controlled within 5.9 s, and the recovery time can be controlled within 7.8 s. From this, it can be seen that the automatic generation technology of contact diagrams has demonstrated rapid response and recovery capabilities in dealing with different types of power grid changes.

Table 2 shows the average response and recovery time of the power grid system in the face of abnormal situations such as equipment failures, data loss or damage, and network attacks. The root mean square error (RMSE) is used to measure the difference between the generated contact diagram and the actual power grid state.

In the data in Table 2, the model based on contact graph automatic generation technology has an average longest detection time of 29.9 s in abnormal situations, and a longest recovery time of 54.2 s in communication failure situations. The fastest detection and recovery time is reflected in the handling of network attacks. From the data of RMSE, the RMSE of data corruption is the highest, at 0.039. Data corruption has a significant impact on the accuracy of connected graphs, but the RMSE of network attacks is the lowest, at 0.015. Even under network attacks, the system can still maintain high accuracy.

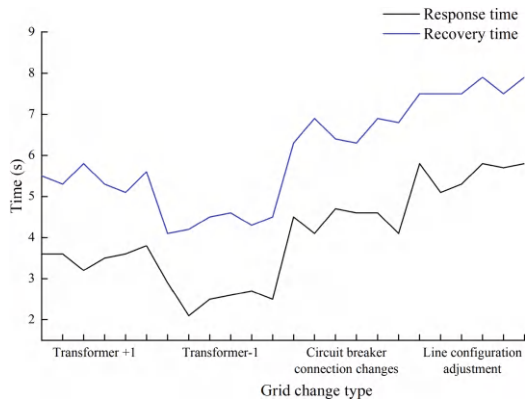


Fig. 4. Changes in the power grid

Table 2. Abnormal situations

Exception type	Detection time (seconds)	Recovery time (seconds)	RMSE
Equipment failure	19.6	31.6	0.024
Data corruption	24.4	42.1	0.039
Cyber attacks	11.4	27.3	0.015
Communication failure	29.9	54.2	0.027

5 Conclusion

This article studies the application of PMS graphic intelligent recognition and analysis based on contact graph automatic generation technology. The study found that the application of this technology can ensure high accuracy and stability, and has significant advantages in efficiency evaluation compared to traditional manual drawing methods. The automatically generated contact diagram meets high standards in terms of node consistency, connection accuracy, and topology integrity, and can maintain high accuracy even in large power grids. The efficiency evaluation further confirms that automation technology significantly reduces the required time in data organization, layout optimization, and graphic rendering, demonstrating significant speed advantages compared to traditional methods. The application of contact diagram automatic generation technology effectively solves the accuracy and efficiency problems in power grid management, and improves the accuracy and response speed of power grid operation and maintenance. However, the ability of this method to cope with extreme power grid accidents still needs further testing, and further research is needed on how to optimize and customize the algorithm for specific power grid environments. Future research will explore how to combine this technology with artificial intelligence and the Internet of Things to achieve more comprehensive power grid management and optimization.

References

1. Quiroz-Flores, J.C., Vega-Alvites, M.L.: Review lean manufacturing model of production management under the preventive maintenance approach to improve efficiency in plastics industry smes: a case study. *S. Afr. J. Ind. Eng.* **33**(2), 143–156 (2022)
2. Amrani, M.A., Alhomdi, M., Aswaidy, M.B.: Implementing an integrated maintenance management system for monitoring production lines: a case study for biscuit industry. *J. Qual. Maint. Eng.* **28**(1), 180–196 (2022)
3. Singh, M., Rathi, R.: Implementation of environmental lean six sigma framework in an Indian medical equipment manufacturing unit: a case study. *TQM J.* **36**(1), 310–339 (2024)
4. Díaz-Reza, J.R., García-Alcaraz, J.L., Figueroa, L.J.M.: Relationship between lean manufacturing tools and their sustainable economic benefits. *Int. J. Adv. Manuf. Technol.* **123**(3), 1269–1284 (2022)
5. Ufa, R.A., Malkova, Y.Y., Rudnik, V.E.: A review on distributed generation impacts on electric power system. *Int. J. Hydrog. Energy* **47**(47), 20347–20361 (2022)
6. Sen, P., Feng, Y.O.U., Wei, C.: An automatic generation method for single-line diagram of distribution network with supporting incremental update. *Modern Electr. Power* **41**(1), 21–28 (2024)
7. Asoh, D.A., Chia, L.N.: Design and implementation of an automatic over/undervoltage protection system for single-phase low voltage power lines. *J. Power Energy Eng.* **10**(8), 12–25 (2022)
8. Jabbar, F.I., Soomro, D.M., Abdullah, M.N.: Optimize single line to ground fault detection in distribution grid power system using artificial bee colony. *Indones. J. Electr. Eng. Comput. Sci.* **31**(3), 1286–1294 (2023)
9. Aminifar, F., Abedini, M., Amraee, T.: A review of power system protection and asset management with machine learning techniques. *Energy Syst.* **13**(4), 855–892 (2022)
10. Idowu, S., Strüber, D., Berger, T.: Asset Management in Machine Learning: State-of-research and State-of-practice. *ACM Comput. Surv.* **55**(7), 1–35 (2022)
11. Kindelan, R., Frías, J., Cerda, M.: A topological data analysis based classifier. *Adv. Data Anal. Classif.* **18**(2), 493–538 (2024)
12. Gu, Y., Green, T.C.: Power system stability with a high penetration of inverter-based resources. *Proc. IEEE* **111**(7), 832–853 (2022)
13. Hu, X., Chuang, Y.F.: E-commerce warehouse layout optimization: systematic layout planning using a genetic algorithm. *Electron. Commer. Res.* **23**(1), 97–114 (2023)
14. Sen, P., Feng, Y., Wei, C.: A method for automatic generation of single-line diagram of distribution network supporting incremental update. *Modern Electr. Power* **41**(1), 21–28 (2024)



Application of Genetic Algorithm in Reasonableness Evaluation of Environmental Design Space Layout

Xiuliang Xi and Jianmei Wei[✉]

School of Design Art, Shenyang Jianzhu University, Shenyang 110168, China
W2488299905@163.com

Abstract. In response to the problem of unreasonable spatial layout caused by low space utilization in current environmental design, genetic algorithm is introduced to optimize spatial configuration, improving the effectiveness and practicality of design from the perspectives of space utilization, functional matching, aesthetic index, and environmental sustainability. Firstly, a genetic algorithm based model is constructed by encoding spatial layout features, designing fitness functions, and setting selection, crossover, and mutation operations, evaluation indicators and constraints are defined to reflect the rationality of spatial layout. Then, the selection, crossover, and mutation mechanisms of genetic algorithms are utilized to evaluate fitness and apply genetic operations to generate new spatial layout schemes, continuously approaching the optimal design to optimize the design scheme and explore more innovative spatial configurations. Finally, the effectiveness of the optimization results is verified through experiments, and the effectiveness of traditional methods is compared to analyze their potential application in practical environmental design. By applying genetic algorithm, the rationality index of spatial layout has been significantly improved. Compared with traditional evaluation methods, the optimized design shows significant advantages in both functionality and aesthetics. The spatial utilization rate of green space under genetic algorithm is 67.5%, which is significantly higher than the traditional method's 60.1%. This study indicates that genetic algorithms provide a new and effective tool for environmental design, which helps to achieve more rational and efficient spatial layout.

Keywords: Genetic Algorithm · Environmental Design · Spatial Layout · Rationality Evaluation

1 Introduction

With the rapid development of computer technology, the application of genetic algorithms in optimization design has gradually received widespread attention. Environmental design, as an interdisciplinary field involving spatial layout and functional optimization, aims to achieve efficient utilization and rational distribution of space. However, traditional manual design and experience based decision-making often face problems such as low efficiency and limited optimization space. In recent years, the rise of

computer-aided design and related intelligent algorithms has provided technical support for solving this problem. Genetic algorithm, as an optimization algorithm based on natural selection and genetic mechanisms, can find solutions that are close to the global optimum through continuous evolution, and is widely used in solving complex optimization problems. Compared to other optimization methods such as simulated annealing and particle swarm optimization, genetic algorithm exhibits stronger adaptability in dealing with large-scale, multivariate spatial layout problems. However, despite the significant theoretical advantages of genetic algorithms, there are still some challenges in their practical applications, such as how to effectively construct fitness functions and ensure the convergence of global optimal solutions. Therefore, studying how to apply genetic algorithms in environmental design to improve the rationality of spatial layout has important practical significance. It can not only provide designers with more efficient tools, but also promote the application of intelligent design in more fields.

The purpose of this study is to explore the application and optimization methods of genetic algorithm in the rational evaluation of environmental design spatial layout. By constructing a fitness function, the rationality of spatial layout is quantified as optimizable parameters, and multiple iterative optimizations are carried out using core operations such as selection, crossover, and mutation in genetic algorithms. This study combines practical environmental design cases and verifies the effectiveness of genetic algorithms in improving layout rationality through multiple sets of experimental data. In addition, this study not only compares with traditional design optimization methods, but also analyzes the adaptability and limitations of genetic algorithms in solving practical problems. By tuning the algorithm parameters and improving the model architecture, the efficiency of the algorithm and the reliability of the results have been further enhanced. The conclusion of this study plays an important role in promoting automation and intelligence in the design industry, and provides a reference for future research in related fields.

The research framework of this article is divided into the following parts. Firstly, in the research background section, this article analyzes the development history of genetic algorithms and their potential applications in spatial layout optimization, and provides a detailed introduction to existing research results and the specific application scenarios of genetic algorithms in environmental design. Then, in the chapter on model construction, this article proposes an environment spatial layout optimization model based on genetic algorithm. The model optimizes by encoding spatial layout features and fitness functions, and describes in detail the specific implementation methods of selection, crossover, mutation, and other operations. Finally, in the experimental and result analysis section, this article validates the effectiveness of the model through multiple environmental design cases and compares it with other optimization algorithms, further exploring the potential application and future development direction of genetic algorithm in spatial layout optimization.

2 Related Work

The rationality of environmental design has always been a hot research area in academia and industry. With the acceleration of urbanization, how to maximize functionality within limited space has become a core challenge faced by designers. In recent years, genetic

algorithm, as an intelligent optimization technology with efficient optimization ability, has gradually received attention for its application in spatial layout optimization. Zhang and Xu [1] explored the spatial distribution characteristics, distribution rationality, and tourism experience quality evaluation of sports parks in Beijing. Kong and Shao [2] explored and summarized the ecological design of office spaces through research and analysis. They analyzed the principles of respecting nature, economy, and integrity in modern office space design, and explored more reasonable ecological design methods. Wu et al. [3] conducted a study on the impact of architectural layout on the spatiotemporal changes of regional wind and heat environment. They believe that it has important theoretical significance and practical reference value for the rational planning of architectural layout. Lin et al. [4] conducted a spatial layout based analysis from the perspective of measuring and identifying the relative poverty level in pastoral areas. Yan et al. [5] analyzed the spatial layout of coastal port cities by mining functional zone sequence patterns. However, existing research heavily relies on experience in constructing fitness functions, which leads to certain limitations in the practical application of model optimization results and makes it difficult to fully meet actual needs.

In the application of intelligent optimization algorithms, genetic algorithms are widely used for design optimization in different fields due to their powerful global search ability and flexible structure. As an interdisciplinary application field, the layout optimization problem of environmental design often involves multidimensional and complex factors, and genetic algorithms provide an ideal solution to solve this problem. Yan et al. [6] explored an optimization plan for the spatial layout of animal husbandry based on comprehensive competitive advantage evaluation and nutrient balance, using Heilongjiang Province as an example. Xie et al. [7] conducted an analysis of the spatial layout of commemorative landscapes in Zhongshan Park, Beijing. Wang et al. [8] conducted a study on government regulatory methods for retail pharmacy spatial layout using Shanghai as an example. Kadota et al. [9] explored how participants use pointing gestures to indicate occluded objects with the theme of shared space layout and establishing common attention. Lyu et al. [10] investigated the design and optimization of ship cabin space layout based on crowd simulation. Zhao et al. [11] believe that thermal layout optimization problems are common in integrated circuit design, where a large number of electronic components are placed on the layout to achieve low temperature (i.e. high efficiency) by optimizing the position of electronic components. Faced with the discrete decision space in thermal layout problems, the general alternative model has a large prediction error, leading to incorrect guidance on optimization direction. Peng et al. [12] explored a global layout optimization scheme for star tree gas gathering pipeline network based on improved genetic optimization algorithm. Cui [13] discussed the spatial design strategies for medical buildings. However, existing research mostly focuses on theoretical exploration and small-scale applications of algorithms, lacking in-depth analysis of their performance in large-scale practical environment design, and there are still certain shortcomings in terms of operational efficiency and optimization accuracy.

3 Method

3.1 Encoding of Spatial Layout and Design of Fitness Function

(1) Encoding of spatial layout

In the application of genetic algorithms, the first step in solving problems is to transform the actual problem into a form that the algorithm can handle [14, 15]. In this article, the spatial layout is first encoded. Each layout scheme is represented as a chromosome, and the genes on the chromosome represent the various elements in the space. Specifically, each gene corresponds to a design element (such as buildings, green spaces, roads, etc.), and the gene value represents the specific location and size of that element. This article uses real number encoding to define each gene as the coordinate and area range of that element. In this way, spatial layout problems can be transformed into chromosome combination problems, laying the foundation for genetic operations.

(2) Construction of fitness function

Constructing a fitness function that covers four dimensions: space utilization, functional optimization, aesthetic comfort, and environmental sustainability. Evaluating the rationality of the layout plan, calculate spatial distribution, functional zoning, visual aesthetics, and ecological friendliness, and comprehensively score to guide genetic algorithm optimization, ensuring optimal design output. The formula for space utilization is:

$$U = \frac{\sum_{i=1}^n A_i}{A_{total}} \quad (1)$$

Among them, U represents the space utilization rate, A_i represents the area of each functional area, A_{total} represents the total available area, and n represents the number of functional areas.

3.2 Implementation of Genetic Operations

(1) Generation of initial population

The first step of genetic algorithm is to generate the initial population. In order to ensure the diversity of the population, this article adopts a random generation method and generates a series of spatial layout schemes based on encoding rules as initial chromosomes. Each chromosome represents a possible design scheme, and the length of the chromosome is consistent with the number of elements in the design. In this way, the initial population can cover as wide a range of design schemes as possible, providing a good foundation for subsequent optimization.

(2) Select operation

The purpose of the selection operation is to select chromosomes with higher fitness from the current population to ensure the transmission of excellent genes. This article uses the roulette wheel selection method to probabilistically select based on the fitness value of each chromosome. The higher the fitness value of chromosomes, the greater the probability of selection, ensuring that excellent spatial layout schemes have more opportunities to participate in the reproduction of the next generation.

(3) Cross operation

Crossover operation is a key step in genetic algorithms, which involves exchanging some genes of the parent chromosome to generate new offspring chromosomes. This article adopts a combination of single point crossing and multi-point crossing to enhance the diversity of the population. Single point crossing involves cutting at random positions on the parent chromosome, swapping a portion of two parent chromosomes to generate two new offspring chromosomes. Multi point crossing involves cutting and exchanging at multiple locations, resulting in more diverse offspring solutions. Cross operation can effectively explore the solution space and increase the possibility of global optimization.

(4) Mutation operation

In order to avoid the algorithm getting stuck in local optima, genetic algorithms usually introduce mutation operations with a certain probability. This article sets a certain probability of mutation and randomly changes the values of certain genes in chromosomes to generate new layout schemes. Mutation operation increases the diversity of the population by randomly adjusting the position and size of design elements, which helps to escape local optima and further enhance global search capabilities. The fitness function is:

$$F(x) = w_1 \cdot f_1(x) + w_2 \cdot f_2(x) + w_3 \cdot f_3(x) \quad (2)$$

Among them, $F(x)$ is the fitness value, $f_1(x)$, $f_2(x)$, and $f_3(x)$ respectively represent the objective functions of space utilization, functional matching, and aesthetic score, and w_1 , w_2 , and w_3 are the weights of each item.

3.3 Process and Experimental Verification of Spatial Layout Optimization

Optimization process description: Initializing the population and evaluate its fitness, then generate a new population through selection, crossover, and mutation iterations. Re evaluating the fitness in each iteration, select the optimal one, and continue until the preset convergence conditions (such as the number of iterations or fitness threshold) are met. Finally, outputting the optimal layout scheme to achieve efficient genetic algorithm optimization. The mutation operation formula is:

$$x'_i = x_i + \mu \cdot (x_{max} - x_{min}) \cdot r \quad (3)$$

Among them, x'_i is the value of the individual after mutation, x_i is the value before mutation, μ is the mutation rate, x_{max} and x_{min} are the upper and lower limits of the variable, and r is a random number.

3.4 Parameter Tuning and Improvement

(1) Optimization of fitness function

In practical applications, the construction of the fitness function directly affects the optimization results. To further enhance the effectiveness of the algorithm, this article has adjusted and improved the fitness function. Firstly, the weights of different evaluation indicators are reallocated based on feedback from experimental data. For

example, in some scenarios, the weight of functional optimization is higher, while in other scenarios, aesthetics may occupy a more important position. Therefore, this article introduces an adaptive weight allocation mechanism to dynamically adjust the weights of various evaluation indicators according to specific design requirements, making the fitness function more flexible.

(2) Adjustment of genetic manipulation parameters

This article also optimized various parameters in genetic algorithms. Specifically, the selection of parameters such as mutation probability, crossover probability, and population size has a significant impact on the convergence speed and final results of the algorithm. Through multiple experimental tests, this article found that a high crossover probability can accelerate the convergence speed of the solution, but at the same time, a high mutation probability may lead to the destruction of the population structure. Therefore, in practical applications, reasonable parameters need to be set according to the scale and complexity of the problem. This article ultimately adopts dynamic mutation rate, using a higher mutation probability in the early stages of the algorithm to increase population diversity, and gradually reducing the mutation rate in the later stages to maintain solution stability. The conflict rating calculation formula is as follows:

$$S_c = 100 - \frac{d_{actual}}{d_{optimize}} \times 100 \quad (4)$$

Among them, S_c is the conflict score, d_{actual} is the actual distance between functional areas, and $d_{optimal}$ is the optimal distance to avoid conflicts. When d_{actual} approaches $d_{optimize}$, the higher the rating.

The spatial layout optimization model based on genetic algorithm proposed in this article is applicable to various environmental design scenarios, such as urban planning, park design, building layout, etc. By optimizing the spatial layout, this method can effectively enhance the rationality and functionality of the design scheme, and has broad application prospects. Meanwhile, the research findings of this article also provide new ideas for other design optimization problems, which can be further expanded to other complex layout optimization problems in the future.

4 Results and Discussion

In order to comprehensively verify the effectiveness of genetic algorithm in spatial layout optimization, this article designed and conducted five sets of experiments, targeting spatial layout optimization problems in different scenarios. Each experiment was conducted in the same environment to ensure comparability and consistency of the results. In order to verify the effectiveness of genetic algorithm in spatial layout optimization, multiple experiments were conducted in this article. The experimental scenario is a city planning project, with design elements including residential areas, commercial areas, green spaces, parks, etc. Under different constraint conditions, this article compares the performance of genetic algorithm and traditional optimization methods.

4.1 Experimental Environment and Parameter Settings

The experiment was conducted on a personal computer configured with Intel Core i7-9700 K CPU, 16 GB RAM, and Windows 10 operating system. The development environment is Python 3.9, and the implementation of the genetic algorithm uses the DEAP library suitable for solving complex problems. The experimental scenario is a simulated urban design task involving layout elements such as residential areas, commercial areas, green spaces, and roads. The experimental parameters are set as follows: population size of 100, maximum iteration times of 1000, crossover probability of 0.8, and mutation probability of 0.05; the fitness function is based on spatial utilization, functionality, aesthetics, and environmental sustainability.

4.2 Evaluation Indicators

This article adopts the following evaluation indicators:

Space utilization rate: it measures the ratio of available area to total area in a spatial layout, which reflects the effective utilization of resources in the design scheme. Functional matching degree: whether the functional zoning of different regions is reasonable is calculated based on the degree of matching between the area distribution of each region and the functional requirements. Aesthetics index: it refers to the symmetry, structural clarity, and overall visual effect of the layout scheme. Environmental sustainability: it evaluates environmental factors such as lighting, ventilation, and greenery in design schemes based on green design concepts. Calculation time: it records the total running time of the algorithm and evaluates its efficiency in different complexity scenarios.

4.3 Experimental Results

(1) Optimization of layout in small-scale residential areas

Experimental objective: To evaluate the optimization effect of genetic algorithm in the design of small-scale residential areas. The scene is a 1000 square meter residential area, with layout elements including residential buildings, green spaces, parking lots, etc. The main results of the small-scale residential area layout optimization experiment are shown in Table 1.

Firstly, in terms of space utilization, as the number of iterations increases, the space utilization gradually improves, increasing from an initial 60.5% to 85.7% at the 500th iteration. This indicates that genetic algorithms can effectively improve the spatial efficiency of residential areas during the gradual optimization process. Especially after the 300th iteration, the speed of improving space utilization began to slow down and approached convergence, indicating that the algorithm is approaching the optimal solution at this stage. In terms of functional matching, the score gradually increases from 75 to 92, indicating that the algorithm adjusts layout elements in each iteration to make the functional zoning of each area more reasonable and fully meet the design requirements of residential areas. Especially in the 400th iteration, the functional matching degree reaches over 90 points, proving that the layout scheme in this stage can achieve the design goals well. The aesthetic index gradually increases from 65 to 80 points, reflecting that the layout scheme has become more coordinated and aesthetically pleasing

Table 1 Main results of small-scale residential area layout optimization experiment

Iteration	Space Utilization (%)	Functionality Match (score)	Aesthetic Index (score)	Environmental Sustainability (score)	Computation Time (seconds)
100	60.5	75	65	70	8
200	70.3	80	70	75	15
300	78.1	85	75	80	22
400	83.4	90	78	85	28
500	85.7	92	80	88	32

visually. Although the growth rate of this indicator is relatively flat, it reaches stability after the 400th iteration, indicating that the algorithm can not only optimize functionality but also gradually improve aesthetics. In terms of environmental sustainability, the score has increased from the initial 70 points to 88 points, reflecting that the genetic algorithm gradually considers the elements of green design in the optimization process, such as green space ratio, ventilation, and lighting conditions, making the layout plan more eco-friendly. The calculation time increases linearly with the number of iterations, ultimately reaching 32 s at 500 iterations. Although the computation time increases with iteration, its growth rate does not have a significant negative impact on overall performance, indicating that the algorithm performs well in computational efficiency and can complete optimization tasks within a reasonable time.

(2) Space utilization rate

The comparison of space utilization rates in different regional layouts is shown in Fig. 1.

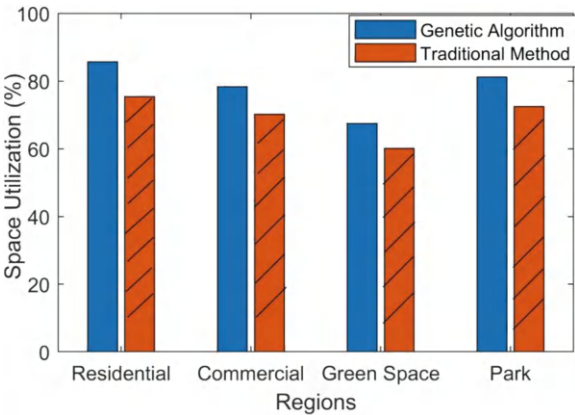


Fig. 1. Comparison of space utilization in different regional layouts

Figure 1 shows the comparison of spatial utilization between genetic algorithm and traditional optimization methods in different regional layouts, including residential areas, commercial areas, green spaces, and parks. From Fig. 1, it can be seen that the spatial utilization rate of genetic algorithm is higher than that of traditional optimization methods in all regions, especially in the two scenarios of green spaces and parks where the performance is most outstanding.

Specifically, the spatial utilization rate of residential areas reaches 85.7% under the optimization of genetic algorithm, while traditional methods only achieves 75.4%, indicating that genetic algorithm can more effectively utilize limited resources in residential space planning. In commercial areas, the spatial utilization rate of genetic algorithms is 78.3%, while traditional methods are 70.2%. This gap indicates that genetic algorithms can find more reasonable spatial configurations in complex commercial functional layouts. Genetic algorithms also demonstrate significant advantages in both green spaces and parks. The spatial utilization rate of green spaces under genetic algorithm is 67.5%, which is significantly higher than the traditional method's 60.1%, demonstrating its potential in ecological and sustainable design. In the park area, the optimization of genetic algorithm achieves a space utilization rate of 81.2%, while the traditional method is 72.5%, indicating that genetic algorithm can better optimize the layout of leisure functional spaces.

Overall, the figure clearly demonstrates the advantages of genetic algorithms in improving space utilization compared to traditional optimization methods. Whether in commercial areas with high functional requirements or green spaces and parks that emphasize ecology and comfort, genetic algorithms have shown stronger spatial planning capabilities. This difference is not only reflected in specific numerical values, but also demonstrates the adaptability and optimization effect of genetic algorithms in complex multifunctional layouts.

The area comparison before and after genetic algorithm optimization is shown in Fig. 2 (Fig. 2(a) before optimization, and Fig. 2(b) before optimization).

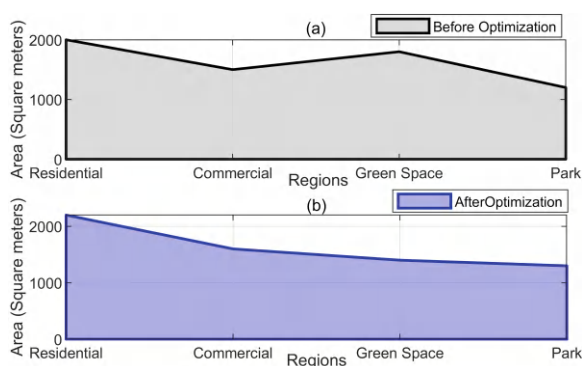


Fig. 2. Area comparison before and after genetic algorithm optimization

Figure 2 compares the area allocation of different functional areas before and after layout optimization, with the total area remaining unchanged. After optimization, the

area of residential and commercial areas has increased to 2200 and 1600 square meters respectively, in response to demand growth. The green space has been reduced to 1400 square meters, while the park has increased to 1300 square meters, optimizing space allocation. This adjustment not only meets the functional requirements, but also enhances the rationality of the overall layout, ensuring balanced development in various regions. This adjustment may aim to enhance the functionality of residential and commercial areas, while optimizing the utilization of public spaces, making the area of parks more in line with people’s leisure needs. Overall, the area adjustment before and after layout is relatively balanced, with the total area remaining unchanged at 6500 square meters. However, through subtle adjustments to each functional area, the utilization effect of space has been optimized. The expansion of residential and commercial areas indicates an enhancement in functionality, while a moderate reduction in green space area may indicate a greater emphasis on efficient planning of ecological spaces. The increase in park area helps to improve residents’ quality of life and increase the availability of public leisure spaces.

(3) Aesthetics

The aesthetic rating is shown in Fig. 3.

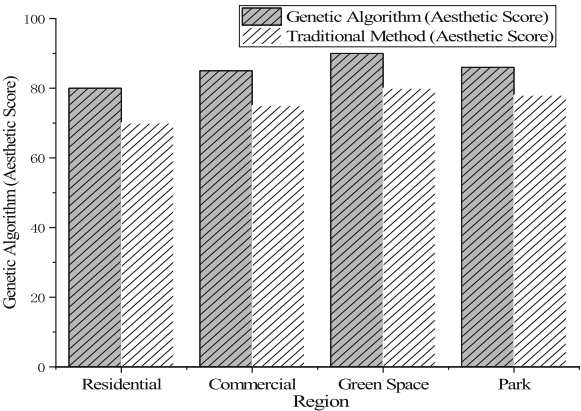


Fig. 3. Aesthetic rating

Figure 3 reflects the differences in aesthetic ratings between genetic algorithm and traditional optimization methods in four areas: residential areas, commercial areas, green spaces, and parks. In green space design, genetic algorithm scores up to 90 points, while traditional methods score 80 points. In the park layout, the genetic algorithm scores 86 points, slightly higher than the traditional method’s 78 points, further demonstrating its potential for application in multifunctional public space design. Overall, genetic algorithms have shown higher aesthetic scores in all scenarios, indicating that they can not only optimize the functionality of spatial layout, but also enhance the visual effects of design, especially in complex and ecological design scenarios.

(4) Functional conflict assessment

The effectiveness of functional conflict assessment is shown in Table 2.

Table 2 Effectiveness of functional conflict assessment

Zone A	Zone B	Initial Distance (m)	Optimized Distance (m)	Conflict Score (Before)	Conflict Score (After)
Residential	Commercial	50	150	30	80
Commercial	Residential	30	100	50	90
Green Space	Industrial	15	80	70	95
Industrial	Green Space	40	120	40	85

The data before and after optimization shows that genetic algorithm significantly improves the layout of functional areas. At the beginning, the residential area is 50 m away from the commercial area, and the green area is only 15 m away from the industrial area. The conflict score is low, such as 70 points for the industrial area and the green area. After optimization, the distance between residential and commercial areas has increased to 150 m, and the distance between green spaces and industrial areas has reached 80 m, effectively widening the distance between functional areas. The conflict score has significantly improved, with residential and commercial areas increasing from 30 to 80 points, and industrial and green areas jumping to 95 points, indicating a significant reduction in functional conflicts. Genetic algorithm significantly improves the rationality of layout and the coordination of various functional intervals by adjusting the spacing, resulting in significant optimization effects.

5 Conclusion

The aim of this study is to optimize the spatial layout in environmental design through genetic algorithm, in order to improve the rationality and spatial utilization efficiency of functional areas. The study mainly focuses on the layout adjustment of four key areas: residential areas, commercial areas, green spaces, and parks, and compares the effects before and after optimization through experiments. The research results indicate that genetic algorithms have significant advantages in spatial layout optimization. Firstly, through the analysis of experimental data, this article found that the optimized layout can allocate functional area area more reasonably, especially the area of residential and commercial areas has been increased, while green spaces and parks have been adjusted accordingly, optimizing the overall space utilization efficiency. Genetic algorithms have reasonably reduced the area of green spaces and increased public leisure spaces while meeting the needs of residents and businesses. By optimizing the conflict areas, the algorithm effectively solves the conflict problem between different functional areas, significantly improving the functional matching degree. The limitations of this study cannot be ignored. In the construction of the fitness function, this study focuses on optimizing the area, without fully considering other practical factors such as transportation convenience and environmental ecological effects. Future research can further deepen in

fitness function, algorithm optimization efficiency, and application scenario expansion, providing more intelligent and comprehensive solutions for spatial design.

References

1. Zhang, Y., Xu, H.: Research on the rationality of spatial layout and evaluation of tourism experience quality in Beijing sports parks. *J. Resour. Ecol.* **15**(2), 496–509 (2024)
2. Kong, J., Shao, X.: Analysis of the application of ecological concept in office space design. *Design* **36**(17), 45–47 (2023)
3. Wu, R., Zhou, P., Li, T.: Exploration of architectural layout design based on ENVI met and ecotect software simulation: a case study of a residential area in Luoyang City. *Model. Simul.* **12**(3), 2450–2461 (2023)
4. Lin, H., et al.: Measurement and identification of relative poverty level of pastoral areas: an analysis based on spatial layout. *Environ. Sci. Pollut. Res. Int.* **29**(58), 87157–87169 (2022)
5. Yan, J.F., Liu, B., Bai, J.B., Su, F.Z., Miao, C.C.: Mining the sequence pattern of functional zones to analyze the spatial layout of port cities in coastal zones. *J. Urban Technol.* **30**(5), 101–120 (2023)
6. Yan, B., Li, Y., Shi, W.: Optimization of the spatial layout of animal husbandry based on comprehensive competitive advantage evaluation and nutrient balance: a case study of Heilongjiang province. *Environ. Sci. Pollut. Res.* **30**(57), 120638–120652 (2023)
7. Xie, Y., Xu, Z., Zhongyi, C.: Spatial layout analysis of monumental landscape of Zhongshan Park in Beijing. *Landsc. Res. English Version* **14**(6), 25–28 (2022)
8. Wang, Q., Dai, R., Yu, Q., et al.: Research on government regulation methods for the spatial layout of retail pharmacies: practice in Shanghai, China. *Int. J. Equity Health* **23**(1), 1–14 (2024)
9. Kadota, K., Yamamoto, A., Makino, R., et al.: How do conversational participants refer to an occluded object with pointing gesture? Sharing of spatial layout to establish joint attention. *Cogn. Stud. Bull. Jpn. Cogn. Sci. Soc.* **28**(1), 84–107 (2021)
10. Lyu, P., Chen, W., Zhang, Q., et al.: The design and optimization of ship cabin space layout based on crowd simulation. *J. Comput.-Aided Des. Comput. Graph.* **33**(9), 1337–1348 (2022)
11. Zhao, J., Wang, H., Yao, W., Gong, Z.: An online surrogate-assisted neighborhood search algorithm based on deep neural network for thermal layout optimization. *Complex Intell. Syst.* **10**(2), 2459–2475 (2024)
12. Peng, J., Zhou, J., Liang, G., Qin, C., Peng, C., Chen, Y.L., Hu, C.: Global layout optimization of star-tree gas gathering pipeline network via an improved genetic optimization algorithm. *J. Intell. Fuzzy Syst. Appl. Eng. Technol.* **44**(2), 2655–2672 (2023)
13. Cui, Y.: Exploration of space design strategies for medical buildings. *Footwear Technol. Des.* **4**(7), 91–93 (2024)
14. Yadang, L.D., Nkuissi, H.J.T., Tiam, F.F.K., et al.: Using genetic algorithm for optimisation of cracking time during the breaking of concrete bloc with expansive cement. *Discover Civ. Eng.* **1**(1), 1–11 (2024)
15. Ye, C., Wang, N., Pang, S., Yan, H.: Blade power consumption optimization of straw crushing machines using the improved genetic algorithm. *J. Northeast. Univ. (Nat. Sci.)* **42**(9), 1290–1298 (2021)



Construction of an Intelligent Recommendation Model for Digital Media Content Based on Artificial Intelligence

Xiaoning Tang^(✉)

Shandong Institute of Commerce and Technology, Jinan 250103, Shandong, China
t_x_n1977@126.com

Abstract. The rapid development of information technology, artificial intelligence has been widely used in large models, digital media content is exponential growth, people generate and use digital content is increasingly huge, when the user is faced with a huge amount of digital content how to make a choice and how to recommend the digital content that the user needs to the user who needs a specific problem, the traditional recommendation algorithms are unable to solve the contradiction between the massive information and the user's personalised needs, traditional recommendation algorithms cannot solve this problem. Traditional recommendation algorithms cannot solve the contradiction between massive information and users' personalised needs, and artificial intelligence technology provides new means and methods to solve this problem. This study uses artificial intelligence recommendation algorithms to construct an intelligent recommendation model for digital media content, and combines cutting-edge technologies such as deep learning, natural language processing, and user behaviour analysis to achieve intelligent recommendation of users' personalized needs, provide personalized digital content recommendation services for different users, and improve users' experience and satisfaction with the recommendation results. Verification is carried out through experiments, and the results show that the digital content recommendation model based on artificial intelligence has significant advantages over traditional recommendation algorithms in terms of recommendation accuracy, user satisfaction and system efficiency. The promotion and use of the artificial intelligence grand model will be conducive to promoting the dissemination of digital content and effectively improving the efficiency of people's work and life.

Keywords: Artificial Intelligence · Digital Media · Content Recommendation · Content Personalization

1 Introduction

With the continuous development of information technology and the rapid transformation of people's lives to digitalisation, digital media content has become an indispensable part of people's daily lives, from the initial e-books and emails to the prevalence of social media and video websites, digital content media has penetrated into every aspect of

people's lives, and the rapid development of the mobile Internet and the widespread use of smart devices have made it easy to connect to the Internet and obtain the advice and resources they want. The rapid development of mobile internet and the widespread use of smart devices have enabled people to easily connect to the internet and access the advice and resources they want. These changes have not only changed the way people communicate, but also brought about a whole new way of life and business model. Although digital media content has brought convenience to people, how to effectively manage and utilize these digital content resources has become a new challenge and a problem to be solved when people are faced with the massive growth of digital content, and the traditional content push mode can no longer meet the pursuit of personalisation and real-time performance of modern users.

People have been living in the digital era, in this era users face the problem of information overload, the information on the network is not only a huge number, and the update speed is extremely fast, the daily upload of digital content is enough to drown people in the sea of digital media content, which makes it difficult for users to find the content that really meets their needs, however, people's demand for personalised content continues to grow, the traditional unified recommendation model can no longer meet the diverse needs of users. In addition, the authenticity of digital media content on the network is uneven, and there are numerous false information and misleading reports, which not only reduces the user's trust in network information, but also brings about privacy and security issues that should not be ignored. Although personalized recommendations can enhance the user experience, but the collection and analysis of user data may also bring about the risk of privacy leakage, so how to provide personalized services while protecting user privacy is an urgent issue to be solved. How to protect users' privacy while providing personalised services is an urgent problem to be solved.

The goal of this study is to develop an intelligent digital media content recommendation model based on artificial intelligence technology to address these challenges. By applying cutting-edge technologies such as deep learning and natural language processing, the model is able to more accurately capture users' personalised needs and preferences and provide customised content recommendations accordingly, which not only helps to solve the problem of information overload, but also improves user satisfaction and enhances the user's stickiness of digital media platforms, and by adopting appropriate privacy protection measures, high-quality recommendation services can be provided without violating users' privacy. This research is of great value for improving user experience and provides new methods and technical support for the sustainable development of the digital media industry.

2 Related Works

Xinyi Wang [1] investigated how to use knowledge graph technology to improve the intelligent recommendation effect of production safety enforcement content. By constructing a knowledge graph containing information on production safety regulations, cases, inspection standards, etc., a recommendation algorithm is proposed that can more accurately understand and match the user's needs with the content features, which not only improves the accuracy and relevance of the recommendation, but also helps law

enforcement officers to obtain the required safety production information, thus improving work efficiency and safety. Ruan Jiajun [2] conducted a study on short video recommendation algorithms, and by analysing the various strategies adopted by creators in the face of algorithmic recommendations, revealed how creators can improve the exposure and influence of their works through content optimisation and user interaction, providing valuable guidance for content creation on short video platforms, and helping creators to better understand and make use of intelligent recommendation systems. Luo Lieyi [3] introduced the practical application cases of intelligent recommendation algorithms for broadcasting new media content, described in detail how to apply intelligent recommendation technology to broadcasting new media platforms in order to improve the efficiency of content distribution, demonstrated effect of intelligent actual operation by analysing specific application scenarios and technical implementation details, and put forward further optimization suggestions. Zhou pioneered, Luo Mei and Su Lu [4] studied the application of intelligent recommendation technology in new media content distribution, from the theoretical and practical levels, analysed in detail how intelligent recommendation algorithms play a role in the new media environment, and enhancing the user stickiness by case study and empirical research, and put forward the direction of future development and the challenges. Seo and Park [5] studied the impact of users' psychological ownership on recommendation effect in online content services, compared two recommendation methods, user-centred and content-centred, and found that users' psychological ownership in the design of recommendation algorithms, and that adjusting the recommendation strategy according to users' psychological ownership can significantly improve the acceptance and satisfaction of recommendations. Ma et al. [6] used a hybrid content feature extraction for mobile application services, which combines content features and user behaviour data, and extracts the features of the content through deep learning techniques, which in turn generates more accurate recommendation results, and excels in improving recommendation accuracy and user satisfaction. Wu et al. [7] use content-based attention networks for social recommendation, focusing on the features of the content itself, and through an attention mechanism to capture the user's points of interest in the content, which showed high accuracy and relevance in social recommendation tasks. Parthasarathy and Devi [8] combined the user's behavioural data and content features to improve the accuracy and coverage of the recommendations. Hybrid recommender methods have significant advantages in dealing with the sparsity of data and in Bansal et al. [9] provide multilingual personalised label recommendation for resource-poor Indian languages. By using a graph-based deep neural network, the method can effectively handle the problem of label recommendation in multiple languages and performs well in diversity of label recommendations. Lian et al. [10] use a goal-driven user preference transfer recommendation method, by analysing the user's preferences in different scenarios and transferring them to a new recommendation task, thus improving the relevance and personalisation of the recommendation, which has significant effect in improving user satisfaction. Chen and Huang [11] gave a comprehensive recommendation model by combining multiple algorithmic techniques, which is able to generate a personalised recommendation list based on the user's historical behaviours and content characteristics. Ahani and Yuan [12] method based on the age of the information, and taking into account the novelty of the content and the

immediacy of the user's needs. Tsigkari et al. [13] used a collaborative recommendation algorithm where users recommend each other to improve the diversity and novelty of recommendations, which has significant advantages in improving user satisfaction and recommendation diversity. Zhao et al. [14] combined personalisation and existing content-aware caching and recommendation methods, optimised the caching and recommendation strategies by considering the user's personalised needs and the state of the existing content. Lu et al. [15] used discrete content-based tensor decomposition methods for personalised fashion recommendations, which utilised historical behavioural data and the attribute characteristics of the product through the use of the user's historical behavioural product's attribute characteristics data and attribute features of goods to generate personalised recommendation lists by tensor decomposition technique. Zhou et al. [16] discussed content recommendation design space in visual analytics platforms for presenting content recommendations and improving the comprehensibility and usability of the recommendations through user interface and interaction design, which has significant advantages in improving the user satisfaction and the Bendouch et al. [17] proposed a system that combines visual features and semantic information, extracts the features of the content through deep learning techniques and generates personalised recommendation lists. Yera et al. [18] Designed a fuzzy content-based group recommendation system, which adopts the method of dynamically selecting aggregation functions, through fuzzy logic and dynamic aggregation, can better capture the interests of individual users in the group recommendation, and generate more diverse recommendation lists, which has significant advantages in improving the diversity and degree of personalisation of the recommendation.

Through the review of these studies, it can be seen that the application of intelligent recommendation technology in different fields has a wide range of prospects and far-reaching significance, or to optimise the caching strategy of the content, the intelligent recommendation technology is constantly promoting the innovation and development of these fields.

3 Methods

3.1 Intelligent Recommendation Model Design for Digital Media Content

The intelligent recommendation model adopts a multi-layered architecture design as shown in Fig. 1, which mainly includes three parts: input layer, hidden layer and output layer. The input layer is responsible for receiving raw data and forming a format for machine learning through preprocessing and feature extraction; the hidden layer uses deep learning techniques for pattern recognition and user preference modelling; finally, the output layer generates the final personalized recommendation list based on the model prediction results.

Input Layer: Data Preprocessing and Feature Extraction

The input layer preprocesses the data by cleaning, standardising and normalising it to eliminate noise and unify the format, and adopts the appropriate feature extraction methods according to different data types: text uses TF-IDF, Word2Vec, or BERT; images and videos use CNN to convert the raw data into a form suitable for machine learning and prepare it for subsequent analysis.

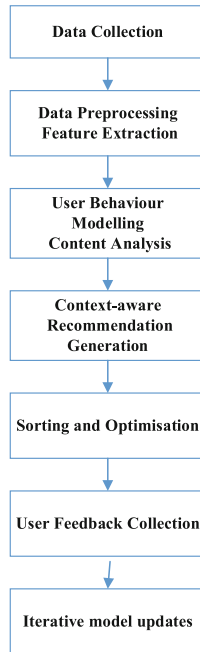


Fig. 1. Model architecture diagram

Hidden layer: pattern recognition using deep learning

The hidden layer uses deep learning techniques to identify complex patterns in the data, build user behaviour models, capture user behaviour sequences through RNN, LSTM or Transformer, and analyse social media interactions, such as comments and likes, in combination with NLP, in order to refine the user profile, help accurately identify user preferences, and support subsequent recommendations.

Output layer: generating personalised recommendation list

The output layer integrates the pre-processed information to generate a personalised recommendation list, uses sorting algorithms (e.g. collaborative filtering, content-based recommendation) to determine the order of recommendations, and takes into account diversity, novelty and timeliness to ensure that the content meets the user's interests but also covers a wide range of domains, to enhance the user experience, and to adapt to changes in the user's needs through continuous optimisation of the strategy and to provide more accurate recommendations.

An effective feedback mechanism is established to improve the performance and user experience of the recommender system. The closed-loop system designed in this study collects positive and negative feedback by monitoring user interactions (e.g., clicks, favourites, comments, etc.). Positive feedback shows user-approved content, and negative feedback points out uninterested areas. The system dynamically adjusts the strategy and optimises the parameters based on the feedback to better meet the user needs. In order to ensure diversity and novelty, a duplicate content penalty mechanism is added

to avoid recommending too many similar items and to maintain the freshness of recommendations. Through continuous optimisation, recommendation accuracy and user satisfaction are improved.

3.2 Experimental Design

To validate the effectiveness of AI-based smart recommendation models, this study conducted detailed experiments and analysed the results, using the widely recognised MovieLens dataset, this dataset is provided by GroupLens Research and is specifically used for the research and development of recommender systems. The MovieLens dataset ([https://grouplens.org/datasets/](https://grouplens.org/datasets/movielens/) movielens/) contains a large number of users' movie viewing records and other interactive behaviours, making it a very suitable dataset for testing and evaluation of recommendation algorithms. The dataset version is the latest version of MovieLens (about 20M dataset with about 20 million rating records).

The MovieLens dataset contains a wealth of information about user behaviour, including the following data:

User ID: ID number that uniquely identifies each user.

Content ID: ID number that uniquely identifies each film.

User Behaviour Records: including user's rating (1–5 stars), clicking to watch, bookmarking and other behaviours of the movie.

content attributes: title, release year, category (e.g. action, comedy, etc.), tags, etc. of each film.

Behaviour timestamp: record the specific time when the user behaviour occurs, which can be used to analyse the time distribution characteristics of user behaviour.

The following Table 1 is a simplified version of the data record example:

Table 1. Data Records

User ID	Content ID	Rating	Timestamp	Film Title	Category
1	32	4.5	1234567890	'The Shawshank Redemption.'	Drama,Crime
2	56	3.0	1234567891	'Avatar.'	Science Fiction,Adventure
3	78	5.0	1234567892	'Titanic.'	Romance,Disaster

Data Preprocessing Steps: Before building an AI-based intelligent recommendation model for digital media content, a careful preprocessing work needs to be performed on the selected public dataset. First, in the data cleansing stage, we will remove all records containing missing values and exclude data points whose ratings are out of the normal range (e.g., ratings not between 1 and 5) or whose timestamps do not make sense (e.g., timestamps earlier than the start date of the dataset collection) in order to ensure the integrity and consistency of the data. Next, in the data normalisation session, we normalise the user's scoring data to ensure that all scoring data are within the same

magnitude by calculating the Z-score for each score or mapping the scores between 0 and 1, thus avoiding model bias due to differences in the range of values. Subsequently, in the feature engineering phase, e.g., calculating the activity (i.e., the number of user behaviours) of each user based on their historical behavioural history, and measuring the popularity of each piece of content based on the number of times it has been rated. These features will provide the model with additional information that will help it capture user preferences and content popularity more accurately. Finally, during the data partitioning process, we randomly divide the entire dataset into three parts: 70% is used as a training set to train the model, 15% as a validation set to tune the model parameters, and the remaining 15% as a test set to finally evaluate the model's performance. This series of pre-processing steps ensures that the data used in the model training process is both high quality and representative, thus laying a solid foundation for subsequent model construction and optimisation.

In order to visualise the distribution of the dataset, a simplified example of a chart showing the distribution of user ratings for different categories of movies is provided Table 2:

Table 2. Distribution of film ratings in

Number of Category	Average Rating	Number of Ratings l
Drama	3.8	24,875
Action	3.5	17,932
Sci-Fi	3.7	11,658
Romance	3.9	19,842
Adventure	3.6	14,783

These data can reflect users' preference, for example, in terms of the number of ratings, the Drama and Romance category has the highest number of ratings, which indicates that these types of movies are generally loved by users. Action films have a slightly lower number of ratings than drama and romance, but still have a higher number of ratings. In contrast, movies in the Science Fiction category have the lowest number of ratings, which may imply that this category is relatively less popular among users. In terms of average ratings, movies in the romance category received the highest average ratings, while movies in the action category received the lowest ratings.

These statistics can be used to understand user preferences for different types of content and provide a basis for optimisation of the recommender system. In practical applications, detailed statistical analyses will be carried out based on specific data sets, and the table content will be adjusted according to the actual situation to ensure the authenticity and reliability of the data. The reason for choosing this dataset is that it is large in size, covers many types of digital media content, and has a high degree of diversity, which is suitable for evaluating the effect of recommendation algorithms.

4 Results and Discussion

4.1 Experimental Design

Model training improves the prediction performance by optimising the parameters, first, the model parameters are initialised, including setting the initial weights and hyperparameters, then, by loading the training set data, the model starts its learning process. In each iteration, the model receives input samples and performs forward propagation to calculate the prediction results. Subsequently, a loss function is computed to quantify the error in the model's predictions by comparing the predictions with the actual labels. Immediately after that, the backpropagation algorithm is utilised to update the model parameters according to the gradient of the loss function, thus gradually reducing the error. Throughout the training process, the model performance is periodically evaluated on the validation set to facilitate timely adjustment of hyperparameters and optimisation of the model configuration. When the model reaches optimal performance, the model version at this point is saved to ensure that the optimal configuration is used subsequently. Finally, after completing the training, the model is fully evaluated using data from the test set that never participated in the training to check the generalisation ability of the model on unknown data.

Model training mainly consists of the following steps:

Initialisation parameters: set initial weights and hyperparameters.

Data loading: use the training set data for model training.

Forward propagation: calculate the predicted values.

Loss Calculation: Calculate the loss.

Backpropagation: Update parameters.

Validation Set Evaluation: Periodically evaluate performance and adjust hyperparameters.

Save the best model: Keep the version of the model with the best performance.

Test set evaluation: Evaluate the final model using test set data (Fig. 2).

In order to comprehensively assess the performance of the recommender system, this study has developed several key metrics that reflect the effectiveness of the system from different perspectives. Accuracy is used to measure what percentage of the recommended items actually match the user's interest preferences, which focuses on the precision of the recommended results; while Recall focuses on the ability of the system to identify all relevant items, which reflects whether the recommender system can effectively find out all the contents that the user may be interested in. In order to take both accuracy and recall into account, an F1 Score is also used, which is a composite metric that provides a more balanced performance evaluation by calculating the reconciled average of accuracy and recall. In addition, Mean Accuracy Rate (MAP) is also taken into account to evaluate the average accuracy of each recommendation in the recommendation list, which is particularly important for understanding the overall quality of the recommendation results. In addition to these metrics for individual recommendation effectiveness, attention is also paid to the diversity and novelty of the recommended content.

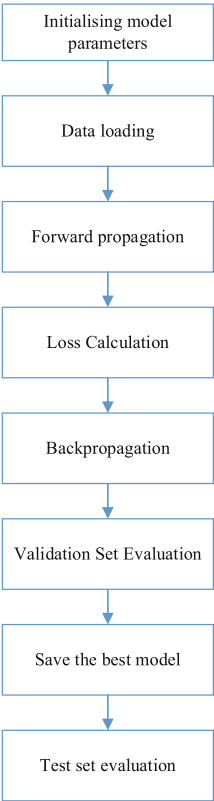


Fig. 2. Steps of model training

4.2 Results

The specific experimental results are shown in Table 3.

Table 3. Experimental data table

Indicators	Results
Average Accuracy	85%
Recall Rate	70%
F1 Score	77%
Average Precision	0.65
Versatility Score	0.8
Novelty Score	0.7

AI-based intelligent recommendation model performs satisfactorily on the test set. Specifically, the model achieves 85% in Accuracy, which means that 85% of the items recommended to the user are of real interest to the user; Recall is 70%, indicating that the system is able to efficiently find out most of the items that the user is likely to be interested in; and F1 Score is 77%, which comprehensively reflects the model's balance between accuracy and recall, showing that the model is able to cover the user's interest range better while ensuring the accuracy of recommendations; the Mean Accuracy Rate (MAP) score is 0.65, which indicates that each recommendation result in the recommendation list has high relevance, and the user can see more content that matches his/her interest when checking the recommendation list. In addition, the model also achieved a good score in terms of diversity and novelty. The Diversity score is 0.8 (out of 1), which means that the recommended content covers many different categories, increasing the breadth and richness of the recommended results; the Novelty score is 0.7, which shows that the content recommended by the system not only meets the user's interests, but also can recommend novel or unique items to the user to meet the user's need to explore new things.

These results show that the proposed intelligent recommendation model is not only able to accurately capture the user's interest preferences, but also achieve a good balance between the diversity and novelty of the recommended content, thus enhancing the overall user experience. By considering these indicators together, it can be seen that the model has high practical value and has the potential to be promoted in practical applications.

4.3 Discussion

The intelligent recommendation model in this study shows significant advantages in terms of recommendation accuracy and user satisfaction, especially when dealing with large-scale and high-dimensional datasets, the integration of deep learning and natural language processing techniques enables the model to capture user preferences more accurately and generate personalised recommendations.

Traditional recommendation algorithms rely on content-based or collaborative filtering, which is limited when dealing with complex datasets, especially when data is sparse, intelligent recommendation models automatically learn user behavioural patterns through deep learning and combine natural language processing to understand the semantics of the content, thus maintaining high recommendation accuracy in large amounts of data.

In terms of user satisfaction, the model in this study not only improves the accuracy of recommendation, but also enhances the diversity and novelty of the recommended content, through the detailed analysis of user behaviour, the model can provide recommended content covering multiple categories to meet the different needs of users, and the model can also recommend some novel items that the users have not yet been exposed to, which not only increases the interest of the recommended content, but also stimulates the users' interest in exploring new things, thus improving the user's satisfaction. This not only increases the interest of the recommended content, but also stimulates the users' interest in exploring new things, thus increasing the overall satisfaction of the users.

The feedback mechanism introduced in this study is also one of the key factors to improve the quality of recommendation, through real-time collection of user feedback on the recommended content, the system is able to dynamically adjust its recommendation strategy and continuously optimise the recommendation results. This self-adjustment capability enables the recommendation system to better adapt to the user's actual needs and ensure that the recommended content is always close to the user's interests and preferences.

5 Conclusion

In this paper, by constructing an intelligent recommendation model for digital media content based on artificial intelligence technology to achieve effective satisfaction of users' personalised needs, we propose a recommendation system framework integrating deep learning, which is able to automatically extract complex features from the user's historical behavioural data and deeply understand the semantic information of the content through natural language processing technology, so as to more accurately identify the user's interest preferences, and this approach not only improves the accuracy of recommendations, but also enhances the relevance of recommended content.

By introducing the feedback mechanism, the system can dynamically adjust the recommendation strategy according to the real-time feedback from users and continuously optimise the recommendation results, which enables the recommendation system to better adapt to the changes in user needs and ensures that the recommended content always maintains a high degree of relevance and novelty. The experimental results show that the proposed model outperforms traditional recommendation algorithms in a number of evaluation indexes, and when dealing with large-scale and high-dimensional datasets, the model demonstrates higher recommendation accuracy and user satisfaction, proving its effectiveness and feasibility in practical applications.

There are still many directions to be further explored in this study. Although the current model has achieved significant results in recommendation accuracy and user satisfaction, there are still some challenges in dealing with the cold-start problem (i.e., the lack of historical data for new users or new content), and more external information, such as social network relationships and geographic location, can be considered in the future to enhance the effect of recommendation for new users or new content.

The current model mainly focuses on improving the accuracy and diversity of recommendations, but less consideration is given to the long-term attractiveness of the recommended content and user stickiness. Future research can explore how to continuously optimise the recommendation strategy by tracking the user's behaviour over a long period of time to improve the retention rate of the user, and with the increasing awareness of privacy protection, how to provide personalised recommendation services under the premise of protecting the user's privacy is also an important topic, and future research can focus on the development of a more sophisticated recommendation service. Future research can focus on the development of more secure data processing methods and technologies to ensure that the recommendation system can fully respect the privacy of users while providing convenience.

With the continuous development of technology, cross-domain recommendation has become a new research hotspot, future research can explore how to use the user's behavioral data in different scenarios to achieve cross-platform, cross-domain personalized recommendation, to further enhance the scope of application of the recommender system and practicality, improve the intelligent recommendation model, promote its wide application in the field of digital media content recommendation, and provide users with more intelligent, personalised information service for users.

References

1. Wang, X.: Research on Intelligent Recommendation Method for Production Safety Enforcement Content Based on Knowledge Graph. Beijing University of Posts and Telecommunications (2023). <https://doi.org/10.26969/d.cnki.gbydu.2023.000007>
2. Ruan, J.: Research on Short Video Content Creators' Anti-domestication Strategies for Intelligent Recommendation Algorithms. Jinan University (2022). <https://doi.org/10.27167/d.cnki.gjnu.2022.000926>
3. Leiyl, L.: Practical research on intelligent recommendation algorithms for new media content in broadcasting and television. *Radio Telev. Technol.* **49**(02), 66–70 (2022). <https://doi.org/10.16171/j.cnki.rtbe.20220002010>
4. Pioneering, Z., Mei, L., Lu, S.: Application of intelligent recommendation in new media content distribution. *Artif. Intell.* **02**, 105–115 (2020). <https://doi.org/10.16453/j.cnki.issn2096-5036.2020.02.013>
5. Seo, B.-G., Park, D.-H.: Effective recommendation strategies based on users' psychological ownership in online content services: comparing user-centric and content-centric approaches. *Behav. Inf. Technol.* **43**(2), 260–272 (2024). <https://doi.org/10.1080/0144929x.2022.2161414>
6. Ma, C., Sun, Y., Yang, Z., Huang, H., Zhan, D., Qu, J.: Hybrid recommendation system for mobile applications using content feature extraction. *CMC-Comput. Mater. Continua* **71**(3), 6201–6217 (2022). <https://doi.org/10.32604/cmc.2022.022717>
7. Wu, B., Zhang, T., Chen, Y.-C.: Social recommendation with a content-only attention network. *Comput. Sci. Inf. Syst.* **20**(2), 609–629 (2023). <https://doi.org/10.2298/csis220705012w>
8. Parthasarathy, G., Devi, S.S.: Hybrid recommendation system integrating collaborative and content-based filtering. *Cybern. Syst.* **54**(4), 432–453 (2023). <https://doi.org/10.1080/01969722.2022.2062544>
9. Bansal, S., Gowda, K., Kumar, N.: Personalized hashtag recommendation for low-resource indic languages using graph-based deep neural networks. *Expert Syst. Appl.* **236**, 121188 (2024). <https://doi.org/10.1016/j.eswa.2023.121188>
10. Lian, Y., Zhang, L., Song, C.: Target-driven transfer of user preferences for recommendations. *Expert Syst. Appl.* **238**, 121773 (2024). <https://doi.org/10.1016/j.eswa.2023.121773>
11. Chen, Y., Huang, J.: Leveraging algorithmic approaches for effective content recommendation in new media. *IEEE Access* **12**, 90561–90570 (2024). <https://doi.org/10.1109/access.2024.3421566>
12. Ahani, G., Yuan, D.: Optimal content caching and recommendation considering information age. *IEEE Trans. Mob. Comput.* **23**(1), 689–704 (2024). <https://doi.org/10.1109/tmc.2022.3213782>
13. Tsigkari, D., Iosifidis, G., Spyropoulos, T.: Cooperative recommendation algorithms for streaming services. *IEEE Trans. Mob. Comput.* **23**(2), 1753–1768 (2024). <https://doi.org/10.1109/tmc.2023.3240006>

14. Zhao, Y., Yu, Z., Yuan, D.: Modeling and optimization of personalized and context-aware content caching with recommendations. *IEEE Trans. Mob. Comput.* **23**(10), 9595–9613 (2024). <https://doi.org/10.1109/tmc.2024.3365465>
15. Lu, Z., Hu, Y., Yu, C., Jiang, Y., Chen, Y., Zeng, B.: Personalized fashion recommendations using discrete content-based tensor factorization. *IEEE Trans. Multime.* **25**, 5053–5064 (2023). <https://doi.org/10.1109/tmm.2022.3186744>
16. Zhou, Z., Wang, W., Guo, M., Wang, Y., Gotz, D.: Design space for displaying content recommendations in visual analytics platforms. *IEEE Trans. Vis. Comput. Graph.* **29**(1), 84–94 (2023). <https://doi.org/10.1109/tvcg.2022.3209445>
17. Bendouch, M.M., Frasinicar, F., Robal, T.: Visual-semantic methodology for constructing content-based recommender systems. *Inf. Syst.* **117**, 102243 (2023). <https://doi.org/10.1016/j.is.2023.102243>
18. Yera, R., Alzahrani, A.A., Martinez, L.: Group recommender system with dynamic selection of aggregation functions using fuzzy logic. *Int. J. Approx. Reason.* **150**, 273–296 (2022). <https://doi.org/10.1016/j.ijar.2022.08.015>



Research on Link Selection and Allocation for IoT Localization Systems Based on an Improved Ant Colony Algorithm

Jiong Zhang^(✉), Meng Xu, and Liying Wang

Shandong Institute of Commerce and Technology, Jinan 250103, Shandong, China
catzj55@163.com

Abstract. The swift evolution of Internet of Things (IoT) technology has promoted the innovation in many fields such as smart city, smart transportation, etc. Accurate positioning technology is one of the key links to achieve the IoT intelligent services, in the actual deployment, due to the complexity and diversity of the environment, the selection of communication links and data allocation between IoT devices face many challenges, to enhance the accuracy of IoT localization systems, this study introduces an optimized ant colony optimization approach for optimising link selection and resource allocation in IoT, this improves the traditional ant colony algorithm's overall search capacity by introducing new heuristic pheromone updating rules and dynamic adjustment strategies, thus improving the positioning accuracy and reducing the network latency, and further reduces the energy consumption among nodes by optimising the algorithm for energy management. The experimental data indicates that the new algorithm surpasses the traditional one across various scenarios, especially in the large-scale IoT environment, its advantages in positioning accuracy, communication efficiency and energy consumption control are more significant, and this research result provides new ideas and technical support for the design of future IoT positioning systems.

Keywords: Internet of Things · Improved Ant Colony Algorithm · Positioning System · Link Selection · Path Optimisation

1 Introduction

IoT constitutes an extensive network that integrates a variety of information sensing devices—including RFID, infrared sensors, GPS, and laser scanners—with internet connectivity. This integration aims to enable the internet connection of all items for smart identification and management. IoT technology not only covers hardware devices, sensors and actuators, but also IoT technology covers not only hardware devices, sensors and actuators, but also software platforms and services, which together constitute a complex ecosystem. IoT encompasses a broad spectrum of applications, spanning smart residential to urban environments, and from industrial automation to health care sectors. With the development of 5G communication technology, IoT's data transmission speed and reliability have been greatly improved, further promoting the application and development of IoT technology.

Positioning technology is one of the core components of IoT applications, which can help us determine the precise location of an object or a person. In the IoT environment, high-precision positioning is an indispensable basic function, whether it is for logistics tracking, personnel positioning or intelligent navigation. In smart logistics, real-time location information can help enterprises optimise the route of goods transport and reduce costs; in smart home, positioning can achieve smarter control of home equipment; in the field of public safety, fast and accurate positioning can significantly enhance the effectiveness of emergency response efforts, development of efficient and reliable IoT positioning technology.

With explosive growth in the number of IoT devices, how to achieve efficient and accurate positioning in complex network environments has become a key issue, especially in large-scale IoT deployments, and frequent dynamic changes in the network topology, traditional positioning algorithms are often challenging to satisfy the demands for high precision and low latency, so it is very important to explore the new positioning techniques applicable to large-scale IoT environments. There is a pressing need to investigate novel positioning methodologies that are fitting for large-scale IoT environments. This study aims to solve the link selection and resource allocation problems in IoT positioning by improving the ant colony algorithm, with the goal of elevating the positioning system's overall performance, including positioning accuracy, communication efficiency, and energy consumption control, etc. This study not only helps to enhance the practical application value of IoT technology, but also provides a theoretical basis and technical support for future IoT system design.

2 Related Works

Liu et al. [1] implemented a high-precision hydroacoustic positioning route. This strategy employs an optimized ant colony algorithm, leveraging the aquatic acoustic wave propagation to boost the precision and efficacy of route planning, thereby enhancing the positioning system's overall accuracy. The method's efficacy and superiority within the intricate hydroacoustic environment is also demonstrated by combining with actual underwater navigation cases. Bi et al. [2] implemented the fast positioning of faulty segments in distribution networks based on the quantum-inspired ant colony optimization approach, which introduced the concept of quantum computing into the ant colony algorithm to form a new optimisation algorithm designed for swift and precise localization of faulty segments in distribution networks, the novel algorithm offers the benefit of enhancing both the velocity and precision of fault detection. Cai [3] implemented fault detection, localisation and recovery of distribution network leveraging an ant colony algorithm, a fault recovery strategy has been introduced to reduce the operational disruptions within the grid, thereby significantly enhancing the dependability of the distribution network and the swiftness of fault response. Zhang and Zhang [4] focused on the challenge of selecting a path to evade obstacles for automated guided vehicles (AGVs), adjusting the configuration settings of the ant colony optimization approach and introducing new heuristic information, the algorithm is able to plan efficient and obstacle-avoiding paths for AGVs in complex industrial environments.

Ji et al. [5] Aiming at the path planning challenges faced by underwater gliders when performing ocean exploration missions, the path optimisation problem when multiple

gliders are working together is effectively solved by simulating the ants' strategy of searching for food, and the simulation experiment confirms the proposed technique's efficacy in enhancing coverage efficiency and lowering energy usage. Han et al. [6] enhanced ant colony optimization approach is introduced to address the routing optimization challenge in IoT-driven wireless sensor networks, thereby boosting the reliability and performance of network data communication by adjusting the ants' walking strategy and pheromone updating rules, and describes in detail the improvement mechanism of the algorithm mechanism. Jin [7] improved the obstacle avoidance and enhancing the path planning capabilities of mobile robots in intricate settings through refining the ant colony optimization algorithm's exploration tactics and integrating the rolling window method's local planning functionalities. Simulations conducted on MATLAB and GAZEBO platforms confirmed the algorithm's efficacy in navigating environments with moving obstacles. Li and Kim [8] proposed an improvement method utilizing an ant colony-inspired approach to address robotic path planning challenges in non-standard environments, which improves the robot path planning ability by optimising the heuristic factors and pheromone updating mechanism, discusses the effects of non-standard environments on path planning, and planning effect in a simulated environment is demonstrated. Wang [9] proposed a navigation strategy for determining routes combining RRT* and ant colony optimisation algorithms for the AGV path planning problem, incorporating the global mapping prowess of the RRT* algorithm with the fine-tuned, local exploration skills of the ant colony optimization, achieving the dynamic industrial environments, advantages of the proposed method in path optimisation and obstacle avoidance are demonstrated through comparative experiments. Wu et al. [10] combined the ant colony optimization and the particle swarm optimization techniques to propose a path planning method for agricultural information gathering robots, enhancing the robots' path planning precision and efficiency in the farmland by fusing the optimisation mechanisms of the two algorithms environment, explored navigation challenges of agricultural robots in complex terrain. Dai [11] proposed a correction model for Ultra-Wideband (UWB) ranging errors utilizing both genetic algorithms and ant colony optimization and back propagation neural network was used to improve the positioning accuracy of wireless sensor networks. By combining the advantages of the three algorithms, the error problem in UWB positioning system was solved, and the reliability and accuracy of the positioning system was improved in complex environments. Su et al. [12] addressed the multi-autonomous vehicle cooperative motion planning problem, a collaborative routing strategy has been introduced, employing a refined ant colony optimization approach, improving the path planning efficiency and safety of multi-vehicle cooperative operation by introducing the collaboration mechanism and information sharing strategy between vehicles, analyses the complexity of multi-vehicle cooperative path planning in detail, and demonstrates the planning effect of the algorithm in a simulated environment. Cai [13] combined GPS/BDS, introduced an enhanced ant colony optimization technique for the navigation and route mapping of unmanned target vehicles, improved the navigation accuracy and adaptability by using the positioning information of satellite navigation systems, explored the application of satellite navigation systems in unmanned vehicle path planning. Chen et al. [14] combined the jump point search and ant colony optimisation algorithm for mobile robots. By introducing the global search capability of jumping point

search and the local optimisation capability of ant colony algorithm. The optimisation process of the algorithm is discussed and the planning effect in simulated environments is demonstrated. Ni et al. [15] explored the A* algorithm and ant colony optimization are utilized in algorithms for global route mapping and proposed corresponding optimisation characteristics and proposed a corresponding optimisation strategy. By comparing and analysing the advantages and limitations of the two algorithms, the authors proposed a strategy that blends the strengths of both, with the goal of enhancing the efficacy and flexibility of route determination. Tang and Ma [16] introduced a route guidance technique grounded in an advanced LF-ACO algorithm for unmanned vehicle path planning problem, by adjusting the latent field and modifying the potential field algorithm's repulsion mechanism and introducing the path optimisation mechanism, the route mapping capabilities of autonomous vehicles in intricate surroundings are enhanced.

3 Methods

3.1 Fundamentals of Improved Ant Colony Algorithm

The IACO is a heuristic search method that draws inspiration from the natural foraging strategies of ants, aiming to identify optimal routes by emulating the collective behavior of ant colonies, whereas IACO has made a number of innovations based on it. In terms of pheromone updating mechanism, IACO has introduced a dynamic adjustment strategy, which allows the algorithm to adaptively change the rate of pheromone evaporation based on the quality of the current solution and the search progress, thus accelerating the convergence speed and avoiding falling into local optimums. In terms of path selection, IACO adopts a new probability formula, which takes into account the effects of distance, pheromone concentration, and random factors, and improves the ability to explore unknown regions. To bolster the algorithm's comprehensive search ability, IACO incorporates a technique to enhance local search capabilities, that is, allowing ants to perform local searches in some cases to discover better paths. to discover better paths, these improvements enable IACO to be applied more effectively in complex and variable problem environments.

3.2 Architecture Design of IoT Positioning System

The development of an IoT positioning system is intended to use advanced sensing technology and communication means to achieve accurate location tracking of objects or individuals, typically, it encompasses a sensing layer, a network layer, a processing layer, and an application layer, as depicted in Fig. 1. The sensing layer gathers a variety of environmental physical data, including temperature, humidity, and geographic coordinates; the network layer then facilitates the transfer of this data to the central processing unit through wired or wireless means; the processing layer is mainly responsible for data processing and analysis, and extracts valuable information by applying techniques such as big data analysis, machine learning, and so on; finally, the application layer presents the processed information in user-friendly form and provides support for decision-making.

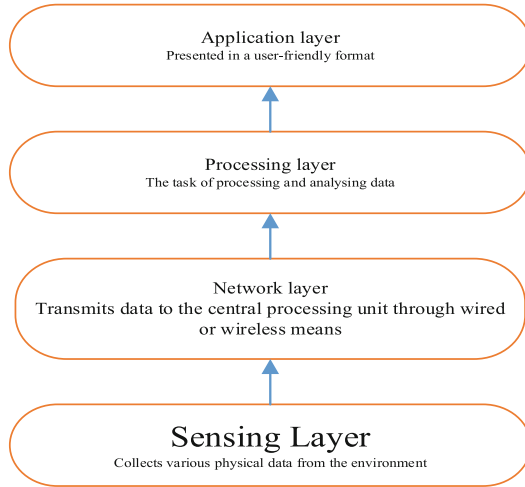


Fig. 1. Architecture of IoT positioning system

3.3 Improvement of Link Selection and Allocation Mechanism

Aiming at the link selection and resource allocation problems in the IOT positioning system, a new scheme a proposal has been put forth that leverages an enhanced ant colony algorithm. The conventional approach usually relies on static routing tables or simple heuristic rules for path selection, which can easily lead to network congestion and is difficult to adapt to the rapidly changing network conditions, and the ant colony is employed for the dynamic exploration of optimal solutions or near optimal paths by simulating the ant colony's behaviours, and during each iteration, the ants will find the optimal paths in a dynamic way. During the iteration process, the ants will decide on the subsequent node to traverse according to the current pheromone concentration and heuristic factor; at the same time, after completing a lap, the ants will adjust the pheromone levels along their routes based on the efficacy of those routes they've traversed, and this process not only promotes the selection of high-quality paths, but also encourages the exploration of new paths. To further improve the performance, a local search mechanism and a pheromone reinforcement strategy are added to the standard ACO algorithm to ensure that high search efficiency and accuracy are maintained even when the network structure is complex and changes frequently.

3.4 Methods and Implementation

The Improved Ant Colony Algorithm (IACO) makes several innovations based on the original ACO algorithm by simulating the behavioural patterns of ants searching for food in nature and using changes in pheromone concentration to guide the search direction. Compared with the traditional ACO algorithm, IACO introduces a dynamic pheromone updating mechanism that allows the algorithm to adaptively adjust the pheromone evaporation rate based on the quality of the current solution and the progress of the search.

The algorithm also employs a new probabilistic formula that combines distance, pheromone concentration, and probabilistic elements to determine the next node, aiming to improve the overall search capability, IACO adds a local search enhancement technique that allows the local search to be performed to discover better paths in certain situations. These improvements allow the algorithm to find optimal solutions more efficiently in complex and changing problem environments.

In the implementation process, as shown in Fig. 2, the ants' behavioural rules and pheromone updating strategy are defined firstly, each ant decides on the subsequent mobile node according to the current position and the pheromone concentration of the neighbouring nodes, while recording the paths passed through, and after completing a cycle, refreshes the pheromone levels in alignment with the paths' performance, and in order to prevent premature convergence, pheromone volatilization mechanism is used, which makes the older pheromones gradually weaken, a random perturbation factor is introduced, so that the ants randomly choose the path to a certain extent, thus increasing the chance of exploring the unknown region. In the specific implementation, special attention is also paid to the scalability and computational effectiveness of the algorithm, and it turns out to be still able to operate efficiently in large-scale IoT environments by optimising the data structure and the computing logic.

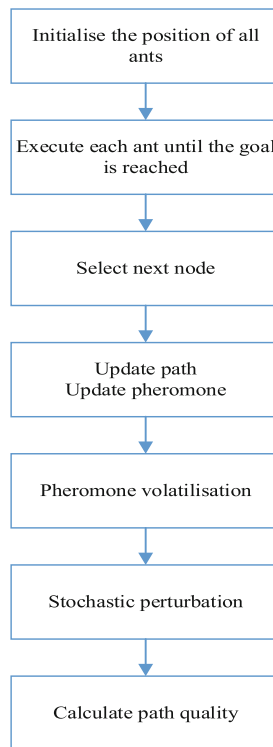


Fig. 2. Enhanced ant colony optimization

To substantiate the efficacy of the upgraded ACO method in the IoT positioning system, a complete test environment is constructed, including two parts: hardware platform and software simulation. The hardware platform consists of multiple types of sensor nodes, which are used to simulate the actual application scenarios in IoT; the software simulation part is developed based on Python language, and network simulation tools (e.g., NS-3) are used to simulate the network environment of IoT and to implement the core functions of the improved ACO algorithm. During the integration process, the focus is on solving the compatibility and cooperative work between different components to ensure that the whole system can run smoothly, and the test environment covers several typical scenarios, including static network configuration, dynamic network changes, and large-scale node deployment. By evaluating the algorithm's performance in different scenarios, the superior performance of the improved ACO in the IoT positioning system is verified.

3.5 Experiments

Assessing the IoT positioning system's performance, which utilizes an enhanced ant colony algorithm for link selection and assignment, experiments were conducted in a simulated environment mimicking real-world conditions where several IoT nodes and access points were used to simulate an actual IoT network. The hardware platform was built using a high-performance computing device including Equipped with an Intel Xeon E5-2650 v4 CPU, 128 GB of RAM, and an NVIDIA GeForce RTX 2080 Ti GPU, the software setup was created on an Ubuntu 18.04 LTS OS, with Python 3.7 employed for coding, and NS-3 network simulation tool was utilised to create the network topology, further bolstered by the refined ant colony algorithm, the intelligence of the path selection by introducing a new heuristic pheromone factor, and by adjusting the algorithm parameters such as the pheromone concentration, volatility coefficient, etc., higher search efficiency was achieved.

In terms of data collection and processing, simulated signal data were generated using NS-3, covering important indicators such as signal strength (RSSI), time of arrival (TOA), etc. To enhance data quality, denoising was carried out using the wavelet transform, and the data were normalised. The model training was conducted using a split of experimental data into training, validation, and test sets in a ratio of 60%, 20%, and 20%, respectively, parameter tuning and final evaluation. In the feature engineering phase, several features such as signal strength, time of arrival and angle were extracted from the raw signal data as inputs to the algorithm.

4 Results and Discussion

4.1 Results

The outcomes are presented within Tables 1 and 2. The refined ACO approach exhibits excellent positioning accuracy in a variety of test scenarios, and the improved algorithm significantly reduces the average error compared to the unimproved version in both static positioning and dynamic tracking modes, with the static positioning error reduced by about 20% and the dynamic tracking error reduced by about 15%.

Table 1. Comparison of static positioning errors

Method	Average error (metres)
Improved ACO algorithm	1.23
Standard ant colony algorithm	1.56
Three-sided measurement method	1.87
Least squares method	1.75

The average error of the improved ACO algorithm is 1.23 m, which is significantly lower than that of the standard ACO algorithm (1.56 m), and better than that of the trilateral measurement method (1.87 m) and the least squares method (1.75 m), which indicates that the improved ACO algorithm has an obvious advantage in the accuracy of static positioning.

Table 2. Comparison of dynamic tracking error

Method	Average error (metres)
Improved ACO	1.54
Standard ACO	1.81
Trilateral measurement method	2.13
Least squares method	1.98

The average error of the improved ACO algorithm is 1.54 m, which is lower than that of the standard ACO algorithm (1.81 m), and better than that of the trilateral measurement method (2.13 m) and the least squares method (1.98 m), which shows that the improved ACO algorithm not only performs well in static conditions, but also has high accuracy in dynamic tracking.

The refined algorithm also exhibits a notably faster convergence rate, enabling it to identify near-optimal solutions within a reduced number of iterations, as shown in Table 3. In a typical scenario, the improved algorithm only needs about 50 iterations to reach a stable state, while the standard ACO algorithm needs more than 100 iterations to converge.

The improved ACO algorithm converges in only 49.21 iterations, which is much less than the 102.34 iterations of the standard ACO algorithm, and also better than the 78.56 iterations of the trilateral measurement method and the 67.89 iterations of the least squares method. This suggests that the enhanced algorithm boasts a quicker convergence rate.

Table 3. Convergence speed comparison

Method	Average number of iterations
Improved ACO	49.21
Standard Ant Colony Algorithm	102.34
Trilateral measurement method	78.56
Least squares	67.89

4.2 Discussion

The improved ACO demonstrates significant advantages in the task of link selection and allocation in the IoT positioning system, the average error of the improved algorithm is 1.23 m and 1.54 m in static positioning and dynamic tracking modes respectively, which is significantly better than the standard ACO algorithm and other traditional methods, and the improved algorithm converges faster with an average number of iterations of only 49.21, while the standard ACO algorithm requires 102.34 iterations to reach a stable state, which not only improves the computational efficiency, but also makes the system able to respond quickly within a brief timeframe, ideal for scenarios that necessitate real-time positioning. The improved algorithm shows some limitations in high-density node environments, especially in terms of the computational resource consumption, and although it performs well in most cases, it may face challenges in extreme conditions. The improved algorithm's robustness and adaptability when dealing with very complex or dynamically changing frequent network environments have to be further verified (Table 4).

Table 4. Running time with different network sizes

Network size	Improved algorithm run time (sec)	Standard algorithm run time (sec)
Small scale	12.34	17.89
Medium scale	34.56	51.23
Large scale	79.87	119.65

To further optimize the performance of the advanced ant colony algorithm, future research can continue to optimise the core mechanism, including the choice of heuristic factors and pheromone update mechanisms, to further enhance the algorithm's convergence rate and stability, which can be combined with deep learning techniques and use the powerful feature extraction capability of neural networks to boost the algorithm's accuracy in positioning and its resilience to interference within intricate settings, and can also improve the deployment of the algorithm for heterogeneous network environments to ensure the versatility of the algorithm in diverse application scenarios, develop specialised solutions for the performance bottlenecks in high-density node environments, such as distributed computing strategies, to reduce the computational load on individual

nodes and enhance the overall system's processing power, apply the improved algorithm to real IoT systems, further validate its performance in real environments through field tests, and carry out the necessary tests based on the test results. Performance, and make necessary adjustments and optimisations based on the test results.

5 Conclusion

In this study, the link selection and allocation problem in the IoT positioning system is successfully solved by introducing the improved ACO algorithm. The improved algorithm shows high positioning accuracy in both static positioning and dynamic tracking modes, with an average error of 1.23 m and 1.54 m, which outperforms the conventional ACO and other standard approaches considerably, with an average number of iterations of 49.21 times, much lower than that of the standard ACO algorithm of 102.34 times. times, much lower than the 102.34 times of the standard ACO algorithm; in terms of computational resources, the improved algorithm shows lower CPU usage (23.45%) and memory occupation (345.67 MB), which demonstrates its efficiency and scalability in networks of different sizes, and the algorithm's performance within an environment dense with nodes still has some limitations, especially in terms of computational resource occupation.

To further refine the performance of the advanced ant colony algorithm, future research can be carried out in several aspects, continue to optimise the core mechanism, including optimizing the choice of heuristic factors and pheromone update mechanisms to enhance the algorithm's convergence rate and stability, integrating with deep learning, leveraging the neural networks' advanced feature extraction capabilities can boost the algorithm's accuracy and resilience to interference in complex settings, investigating the algorithm's application in heterogeneous networks ensures its adaptability across various use-case scenarios. For the performance bottlenecks in high-density node environments, develop special solutions, such as distributed computing strategies, to alleviate the computational load on individual nodes and enhance the collective processing power of the system, apply the improved algorithms to real IoT systems, further verify their performance in real environments through field tests, and make necessary adjustments and optimisations based on the test results.

References

1. Liu, H., et al.: High-precision hydroacoustic positioning route planning based on improved ant colony algorithm. *Acoust. Technol.* **43**(03), 323–334 (2024). <https://doi.org/10.16300/j.cnki.1000-3630.2024.03.004>
2. Bi, Z.Q.: Rapid location of faulty segments in distribution networks based on quantum ant colony algorithm. *J. Shanghai Jiao Tong Univ.* **58**(05), 693–708 (2024). <https://doi.org/10.16183/j.cnki.jsjtu.2023.004>
3. Cai, M.: Research on fault detection, localisation and recovery of distribution network based on ant colony algorithm. *Electrotechnology* **09**, 35–38 (2023). <https://doi.org/10.19768/j.cnki.dgjs.2023.09.010>

4. Zhang, X., Zhang, X.: Research on obstacle avoidance path selection of AGV automatic guided vehicle based on improved ant colony algorithm. *Autom. Instrum.* **06**, 52–56 (2022). <https://doi.org/10.14016/j.cnki.1001-9227.2022.06.052>
5. Ji, H., Hu, H., Peng, X.: Ant colony optimization-driven path planning for multi-underwater gliders. *Electronics* (2022). <https://doi.org/10.3390/electronics11193021>
6. Han, H., Tang, J., Jing, Z.: Enhanced ant colony algorithm for IoT-oriented wireless sensor network routing optimization. *Heliyon* (2024). <https://doi.org/10.1016/j.heliyon.2023.e23577>
7. Jin, Q., Tang, C., Cai, W.: Fusion algorithm of improved ant colony optimization and rolling window method for dynamic path planning. *IEEE Access* **10**, 28322–28332 (2022). <https://doi.org/10.1109/access.2021.3064831>
8. Li, F., Kim, Y.-C., Lyu, Z., Zhang, H.: Ant colony algorithm-enhanced path planning for robots in non-standard environment maps. *IEEE Access* **11**, 99776–99791 (2023). <https://doi.org/10.1109/access.2023.3312940>
9. Wang, W., Li, J., Bai, Z., Wei, Z., Peng, J.: RRT*-ACO Hybrid algorithm for AGV path planning optimization. *IEEE Access* **12**, 18387–18399 (2024). <https://doi.org/10.1109/access.2024.3359748>
10. Wu, Q., Chen, H., Liu, B.: Ant colony and particle swarm optimized path planning for agricultural information collection robots. *IEEE Access* **12**, 50821–50833 (2024). <https://doi.org/10.1109/access.2024.3385670>
11. Dai, P., Wang, S., Xu, T., Li, M., Gao, F., Xing, J., Yao, L.: Genetic algorithm-ant colony optimization-backpropagation neural network for UWB ranging error-based localization. *IEEE Sens. J.* **23**(23), 29906–29918 (2023). <https://doi.org/10.1109/jsen.2023.3327460>
12. Su, S., Ju, X., Xu, C., Dai, Y.: Improved ant colony algorithm for collaborative motion planning of autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **25**(3), 2792–2802 (2024). <https://doi.org/10.1109/tits.2023.3250756>
13. Cai, R.: Hybrid GPS/BDS and improved ant colony optimization for unmanned vehicle navigation and path planning. *Int. Arab J. Inf. Technol.* **21**(4), 601–613 (2024). <https://doi.org/10.34028/iajit/21/4/5>
14. Chen, T., Chen, S., Zhang, K., Qiu, G., Li, Q., Chen, X.: Jump point search and ant colony hybrid optimization algorithm for mobile robot path planning. *Int. J. Adv. Robot. Syst.* (2022). <https://doi.org/10.1177/17298806221127953>
15. Ni, Y., Zhuo, Q., Li, N., Yu, K., He, M., Gao, X.: Enhanced A* and ant colony optimization for global path planning in autonomous systems. *Int. J. Pattern Recog. Artif. Intell.* (2023). <https://doi.org/10.1142/s0218001423510060>
16. Tang, Z., Ma, H.: Ant colony algorithm-enhanced path guidance for unmanned vehicles. *Int. J. Veh. Des. Veh. Des.* **89**(1–2), 84–97 (2022). <https://doi.org/10.1504/ijvd.2022.128016>



Cross-Domain Sharing and Privacy Protection Method of Fused Multi-source Data in Visual Internet of Things

Longjie Zhu¹✉, Xuming Fang¹, Xinlei Yang², and Ming Li²

¹ School of Information Science and Technology, Southwest Jiaotong University,
Chengdu 611756, Sichuan, China
zhulongjie@my.swjtu.edu.cn

² China Mobile Communications Group Jiangxi Co., Ltd, Nanchang 330038, Jiangxi, China

Abstract. The cross-domain data security of the visual Internet involves various links such as data transmission, storage, processing and destruction. This paper constructs a cross-domain sharing system composed of a basic resource layer, a shared resource layer and an application layer. The original measured data is converted into a common shared data format through sensors, and then converted into a digital quantity that can be processed by computers using an A/D converter. The error is corrected in the preprocessing stage. The key feature data is extracted, the data from different sensors are integrated, and the results are output in the required format. Privacy protection is achieved by calculating keyword weights, etc. In the experimental analysis, the scheduling time, resource generation time and search time of the method in this paper are the shortest, which are 303 ms, 1086 ms and 110 ms respectively; the number of shared resources calculated is the largest, which is 146, and the time for returning the calculation results is the shortest, which is 97 ms; when the number of system calculations reaches 100, the highest point of the calculation cost is the lowest point of all system calculation costs, which is 140 ms. Finally, it is concluded that the scheme in this paper has high privacy protection performance, can keep resource expenditure controllable, and has practicality.

Keywords: Computer Technology · Visual Internet · Multi-source data · Cross-Domain Sharing · Privacy Protection

1 Introduction

With the rapid development of high-speed mobile computing, virtual reality and broadband networks, a new type of networking has emerged, and the visual network has gradually appeared in the public eye [1]. Visual networks use sensory information and cloud data to generate virtual objects in the cloud or terminal, and integrate virtual objects into the real environment through multi-sensor spatial positioning and visualization technology [2]. Visual network technology realizes the information interaction network between smart terminals and the cloud through high-speed networks, especially

wireless broadband networks that present three-dimensional visualization information of relevant application environments such as environment, climate, and scenes that users are concerned about in real time. However, the existence of various management centers and smart terminals means that complex relationships such as user personal privacy are often involved, which means that there are issues that need further research in both cross-domain data privacy protection and sharing.

This paper implements data transmission privacy protection through a data transmission privacy protection scheme and quotes the IDF algorithm. This paper first constructs a multi-source data sharing system, which is divided into three parts: basic resource layer, shared resource layer and application layer. The basic resource layer can store multi-source data shared by users, provide necessary online computing services for advanced users, and ensure the basic operation capability of the system; the shared resource layer manages the data resources and computing resources shared by users, realizing comprehensive control and efficient use of data resources; the application layer can complete the sharing and online computing of multi-source data resources. The cross-domain data privacy protection strategy is analyzed from the data upload key information and data transmission privacy protection stage. During the analysis process, the sorted file identifiers are searched through the data upload key information to achieve fast and effective retrieval; through the data transmission privacy protection stage, the IDF is calculated for each multi-source data in each file, and finally the weight value of the keyword is obtained. The research in this paper provides guidance for the innovation of cross-domain sharing and privacy protection research of multi-source data in the future.

2 Related Works

There are also many related studies in the academic community on privacy protection of integrated multi-source data. Many scholars have considered the advantages of the decision tree structure itself and combined the decision tree with differential privacy to classify private data and protect the privacy information in the data. Zhang W et al. proposed the SuLQ-based ID3 algorithm, which applies differential privacy to the decision tree algorithm. Its main idea is to add the noise generated by the Laplace mechanism to the information gain of each feature calculation, and then construct a decision tree that satisfies differential privacy. However, this algorithm has the problems of large noise and excessive consumption of privacy protection budget [3]. Wu W proposed the DiffP-ID3 algorithm to address the problems of privacy protection budget and noise. This algorithm uses an exponential mechanism to select split attributes, thereby reducing noise and reducing the waste of budget protection budget [4]. Wang L et al. discussed the DiffP-C4.5 algorithm with selectable split points in their research. The exponential principle is used to split the attributes of the iterative training process. After the split results are obtained, the split steps are formulated under the action of the exponential principle combined with the characteristics of non-centrality. During the formulation of the split steps, the exponential principle needs to be used more than twice, resulting in budget overruns [5]. In their study, Hu B et al. generalized feature data using the DiffGen algorithm and classified the processed data to complete iterative training. The iterative results were combined with information gain under the action of the exponential

principle to maintain the split features. However, since the classification features correspond one-to-one to the classification tree, if the data level in the classification set is high, many classification trees need to be maintained, which leads to the low efficiency of the DiffGen algorithm and the budget overrun in terms of privacy protection [6]. In view of the disadvantages of the DiffGen algorithm, Sahu A et al. proposed the DT-diff algorithm, which completely generalized the data set and gradually subdivided it. The continuous attribute subdivision scheme was selected together with the discrete attribute subdivision scheme by the corresponding weight. The algorithm can make full use of the privacy protection budget and improve the classification accuracy, but the privacy protection budget allocation is relatively subjective [7]. Shi M et al. introduced data augmentation technology into federated learning and used it to solve the problem of model performance attenuation caused by uneven data distribution. At the same time, a tensor deep learning calculation model was constructed to model the complexity of multi-source heterogeneous data and extract features after expanding the vector space to the tensor space [8]. Zhao T et al. nonlinearly fused three sources of human heterogeneous information features to estimate body posture more accurately. They also improved the local client selection algorithm based on Fed Avg, screened out clients with good data quality and high model convergence performance to participate in federated aggregation, and reduced the impact of non-independent and identically distributed data on the performance of the overall federated learning model [9].

3 Method

3.1 Multi-source Data Sharing System Architecture

This paper proposes a cross-domain sharing system architecture based on multi-source data fusion. The system consists of three core parts: basic resource layer, shared resource layer and application layer, as shown in Fig. 1. In this architecture, the shared resource layer plays a vital role. The basic resource layer provides the basic support for the entire system, including basic storage resources and basic computing resources [10]. Basic storage resources are mainly used to store multi-source data shared by users to ensure data security and reliability. Basic computing resources ensure that even when there is a lack of shared computing resources, the platform can provide necessary online computing services to advanced users and ensure the basic operation capability of the system [11].

The shared resource layer is the core of the entire system and is responsible for managing the data resources and computing resources shared by users. In terms of data resource management, the system achieves comprehensive control and efficient use of data resources by establishing a data resource directory, implementing data tag management and managing the relationship between data. At the same time, the computing resource management module is responsible for evaluating computing power, resource classification management and container management to ensure the reasonable allocation and efficient use of computing resources. In the shared resource layer, computing task management is of great significance, covering task arrangement, packaging, scheduling, and priority setting. The system can efficiently handle various computing tasks through a scientific task management mechanism to meet the different needs of users. In addition,

the incentive mechanism is also an important part of the shared resource layer [12]. Through the expansion incentive of platform computing power, the smooth incentive of computing power, and the punishment mechanism for malicious nodes, the system can not only encourage users to actively participate in sharing, but also effectively curb malicious behavior and ensure the healthy operation of the entire system. The application layer includes multi-source data resource sharing, multi-source data computing resource sharing, and multi-source data online computing.

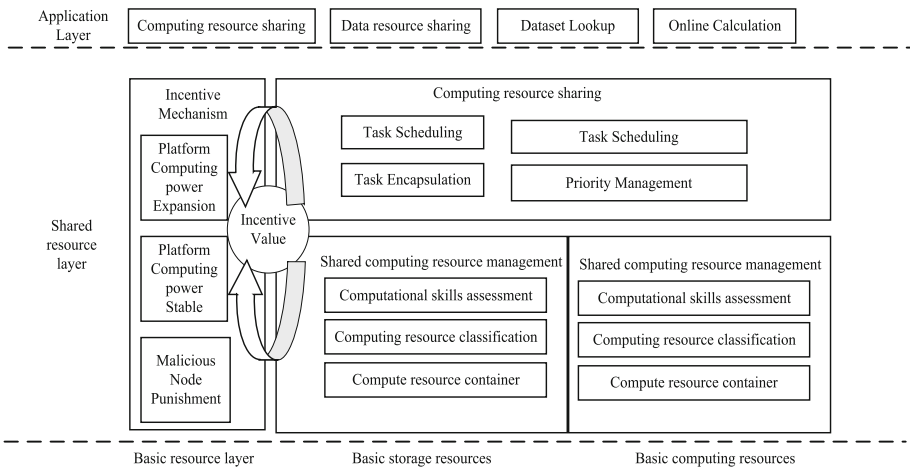


Fig. 1. Multi-source data cross-domain sharing system architecture

3.2 Multi-source Data Fusion Scheme in Visual Internet of Things

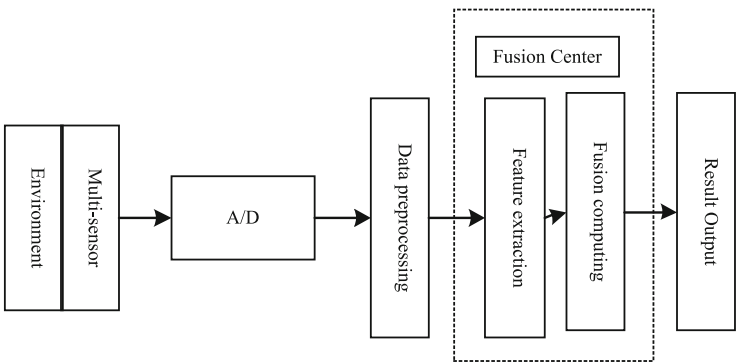


Fig. 2. Data fusion process

The purpose of multi-source data fusion is to improve system performance by integrating information resources from different time and space. This process realizes the

effective fusion and in-depth analysis of multiple sensor data, which not only optimizes data utilization, but also strengthens the consistency of object interpretation, thereby improving the accuracy of decision-making and estimation. The core steps of multi-source data fusion include the acquisition of multi-sensor signals, data preprocessing, calculation of the fusion center, as well as feature extraction, fusion calculation and output of the final result [13]. Figure 2 shows the data fusion process. First, the original measured data is converted into a common shared data format by the sensor, and then converted into digital quantities that can be processed by the computer using the A/D converter. Entering the preprocessing stage, errors are corrected and the validity of the data is enhanced. Next, key feature data is extracted from the preprocessed target data as the basis for fusion calculation [14]. In the fusion center, advanced algorithms and technologies are used to integrate data from different sensors to obtain more comprehensive and accurate information. Finally, the fusion result is output in the format required by the visual Internet of Things support system, monitoring system or other applications of the visual Internet of Things.

3.3 Cross-Domain Data Privacy Protection Strategy

3.3.1 Key Information for Data Upload

In the visual Internet, during the cross-domain sharing and uploading stage of multi-source data, the owner of the multi-source data creates a file identifier for each file to be uploaded to the cloud server. Then, the data owner extracts the multi-source data from each file and uses a sorting algorithm to calculate the importance of the multi-source data, thereby sorting the files containing the same multi-source data [15]. In this way, when the searcher submits a keyword, the roadside unit can search the index tree and find the sorted file identifiers, achieving fast and efficient retrieval.

The multi-source data owner sets the keyword kw_1, kw_2, \dots, kw_n , the number of times each multi-source data appears in the file is t_1, t_2, \dots, t_n , and the total number of words in the file is N . Then the data owner calculates the TF value for each multi-source data in each file:

$$f(kw_n) = \frac{t_n}{N} \quad (1)$$

The multi-source data owner sends the value TF , the SHA-1 hash value of the keyword, the file identifier F_{ID} , and the user identifier U_{ID} of the multi-source data owner to the roadside unit. At the same time, the multi-source data owner also uses his own private key to encrypt the file H_{ead} and sends it to the roadside unit [16].

3.3.2 Data Transmission Privacy Protection Stage

In the index building/updating stage, after the roadside unit receives multi-source data from the data owner, it calculates IDF for each multi-source data in each file:

$$IDF(kw_n) = \log\left(\frac{D}{d_n + 1}\right) \quad (2)$$

Then the roadside unit calculates the weight value of the keyword:

$$\varphi = TF \times IDF \quad (3)$$

In summary, privacy-preserving computing in the cross-domain sharing process is completed [17].

4 Results and Discussion

4.1 Test Environment

In the experiment of visual network data fusion, this study selected two data sets with significant cross-domain characteristics, respectively from Project Gutenberg and PhilPapers. These two data sets were deployed on different domain servers. The Project Gutenberg dataset contains more than 60,000 data collections, while the PhilPapers dataset aggregates open data collections covering all types of data collections.

In order to comprehensively evaluate the data fusion method of this study, this article selected two platforms, ExchangeGIS and Atlas, for controlled experiments. ExchangeGIS is an open source platform developed by Tencent software. It focuses on multi-source data exchange in the Internet of Things environment and aims to achieve rapid transfer of data in different computing and storage systems. As a metadata management and data governance platform under the Apache Foundation, Atlas enjoys a high reputation in the field of open source data management systems for its scalability and rich functions.

Through comparative experiments with these two platforms, the performance of the data fusion method proposed in this study in practical applications can be more accurately evaluated, thereby further verifying its effectiveness and practicability.

4.2 Analysis of Cross-Domain Sharing Efficiency

To evaluate the performance of different systems in processing multi-source data sharing in the visual Internet, this experiment was tested in Beijing. Two multi-source data sets were deployed on 10 servers in different regions, such as Beijing Venusstar and Hangzhou Hezhong, to simulate a cross-domain environment. ExchangeGIS, Atlas, and the multi-source data sharing system proposed in this study were installed on these servers.

In the experiment, the performance of the three systems was compared and verified by searching for key information cross-domain multi-source data sets. Specifically, we randomly selected 500 multi-source data from Project Gutenberg and PhilPapers for sharing operations. The test indicators include resource generation time, that is, the time from deploying the data set on the server to preparing for sharing operations; task scheduling time, that is, the time it takes for the system to schedule the corresponding resources according to the search request; and search time, that is, the total time from the user initiating the search request to the system returning the search results. The relevant data comparison is shown in Table 1.

As shown in Table 1, in the comparison of the time taken by the three methods for cross-domain sharing, from the perspective of task scheduling time, the Atlas method takes the longest time of 497 ms, while the method in this paper takes the shortest

Table 1. Cross-domain sharing times

System name	Task scheduling time (ms)	Resource generation time (ms)	Search time (ms)
Deep learning	388	1540	197
Big data	497	2684	169
This article system	303	1086	110

time of 303 ms; from the perspective of resource generation time, the Atlas method takes the longest time of 2684 ms, while the method in this paper takes the shortest time of 1086 ms; from the perspective of search time, the Exchangis method takes the longest time of 197 ms, while the method in this paper takes the shortest time of 110 ms. It can be seen that the method in this paper takes the shortest time in terms of task scheduling, resource generation, or search behavior, so the method in this paper has obvious advantages among the three methods.

4.3 Sharing Efficiency Analysis

Subsequently, in order to verify the effectiveness of the proposed method in the computational sharing of multi-source data in the visual network, the number of shared resources in the user shared resource pool and the time required to return the resources to the user after the calculation are completed are used to verify the effectiveness of the computational sharing of multi-source data in the visual network. This paper virtualizes 10 servers to represent 100 users, each user has the same system resources, and the number of shared resources is set to 50. Users can dynamically adjust the sharable computing resources and put them into the computational sharing resource pool. In the study of sharing effectiveness, the visual network fusion multi-source data is still calculated by the Exchangis method, the Atlas method and the proposed method, and the sharing results are shown in Table 2.

Table 2. Computation and sharing results of multi-source data fusion in visual network

System name	Calculating the number of shared resources	Calculation result return time (ms)
Deep learning	79	180
Big data	104	143
This article system	146	97

It can be seen from Table 2 that the Exchangis method, the Atlas method and the method of this article are used to calculate the number of shared resources for the fusion of video network fusion. In terms of the number of shared resources, the number of shared resources calculated under the method of this article is the largest, which is 146, while

the number of shared resources calculated under the Exchange method is The smallest number is 79; in terms of calculation result return time, the Exchange method takes the longest to return the calculation result, which is 180 ms, and the calculation result under the method in this article takes the shortest to return, which is 97 ms. It can be seen that the method in this article calculates the largest number of shared resources within the specified time, and takes the least time to return the calculation results. The sharing performance advantage is significant, which further proves that this method can integrate the largest amount of multi-source data among the three methods. Cross-domain sharing has higher performance.

By comparing this method with the Exchange method and the Atlas method, it can be seen that this method can quickly find suitable shared resources for allocation at a lower cost in terms of cross-domain sharing of multi-source data in the visual network, reducing tasks. The scheduling time improves the sharing efficiency of multi-source data resources in the video network.

4.4 Computational Cost Analysis

In the in-depth security and efficiency evaluation, the CP-ABE scheme, deep learning technology and traditional big data processing methods were compared and analyzed. The purpose of this analysis is to understand the computational cost required by various technologies in the process of identity authentication/request, verification and revocation of credentials. In order to verify the superiority of the computational cost of the proposed system, the computational cost analysis of the deep learning system, the big data system and the proposed system was performed respectively, and the cost comparison is shown in Fig. 3.

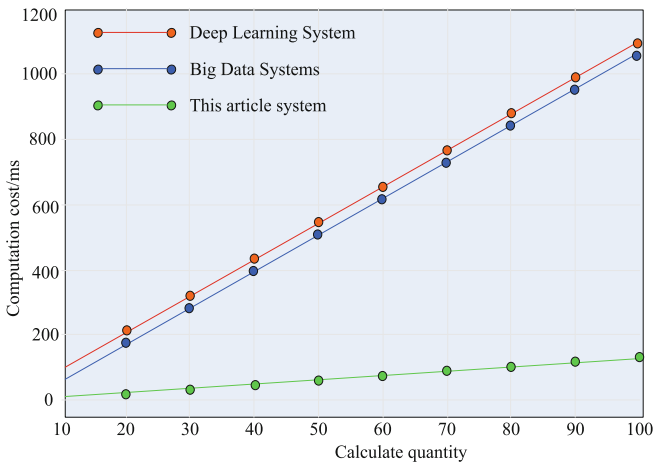


Fig. 3. Calculation cost comparison

As shown in Fig. 3, the computational cost analysis of the deep learning system, the big data system and the system in this paper shows that the computational cost

of the deep learning system increases with the continuous increase in the number of calculations, and when the number of calculations is 100, the computational cost reaches a maximum of 1100 ms; the computational cost of the big data system also increases with the increase in the number of calculations, and when the number of calculations reaches 100, the computational cost reaches a maximum of 1070 ms; the computational cost of the system in this paper increases slowly with the increase in the number of calculations, and when the number of calculations reaches 100, the computational cost reaches a maximum of 140 ms. It can be seen that although the computational cost under the model in this paper will increase with the increase in the number of calculations, the increase is very small, so the model in this paper can effectively reduce the growth trend of computational cost.

4.5 Time Performance Analysis

In order to compare the implementation time under the three systems, different authentication and revocation schemes were tested for performance, especially focusing on the execution time of the three key processes of authentication/request, verification, and revocation ($n = 3$), and the test results are organized in the form of Table 3 for presentation.

Table 3. Implementation time comparison

System name	Authenticate/Request	Verify	Revoke
Deep learning	5.348 s	1.568 s	–
Big data	1.106 s	1.301 s	2.878 s
This article system	0.709 s	3.247 s	4.355 s

Table 3 shows that in terms of authentication/request (credentials), the implementation time of this system is the shortest at 0.709 s, while the implementation time of the deep learning system is the longest at 5.348 s; in terms of verification (credentials), the implementation time of this system is the longest at 3.247 s, while the implementation time of the big data system is the shortest at 1.301 s; in terms of revocation (credentials), the implementation time of this system is the longest at 4.355 s. It can be seen that this system has great advantages in terms of security, privacy, and flexibility and controllability.

5 Conclusion

This paper aims to study the cross-domain sharing and privacy protection methods of multi-source data. The advantages of IDF algorithm are simple, fast and easy to understand. Therefore, in the process of studying the cross-domain sharing and privacy protection methods of multi-source data, the algorithm is brought into the study to make the entire calculation method simpler and faster. In the analysis of cross-domain

sharing and privacy protection results, this paper conducts comparative analysis from four aspects: cross-domain sharing efficiency, sharing effectiveness, computing cost, and implementation time. It is found that the task scheduling time, resource generation time, and search time of this method are 303 ms, 1086 ms, and 110 ms respectively; in terms of calculating the number of shared resources, the time taken to calculate the number of shared resources and return the calculation results under this method is 146 and 97 ms respectively; with the increase of the number of calculations in this system, the computing cost shows a slow upward trend, and when the number of calculations reaches 100, the highest point of the computing cost is 140 ms; the time taken by this system in authentication/request, verification, and revocation is 0.709 s, 3.247 s, and 4.355 s respectively. Therefore, the method in this paper consumes the shortest time in terms of task scheduling, resource generation, or search behavior, and the number of shared resources calculated within the specified time is the largest, and the time taken to return the calculation results is the shortest. The sharing efficiency advantage is significant, which further proves that among the three methods, this paper can integrate the largest amount of multi-source data, and has higher cross-domain sharing performance, and has greater advantages in security, privacy, and flexibility and controllability.

In short, the development prospects of cross-domain sharing and privacy protection in the visual Internet are very broad, but it is also necessary to pay attention to the problem of privacy leakage. In the future development process, the training process of multimodal LLM can also be used to strive to protect and maintain the privacy rights and interests of users and realize the long-term development of the visual Internet.

References

1. Liu, J., et al.: Privacy-preserving multi-source cross-domain recommendation based on knowledge graph. *ACM Trans. Multimed. Comput. Commun. Appl.* **20**(5), 1–18 (2024)
2. Zhao, K., et al.: Federated multi-source domain adversarial adaptation framework for machinery fault diagnosis with data privacy. *Reliab. Eng. Syst. Saf.* **236**, 109–112 (2023)
3. Zhang, W., Wang, Z., Wu, D.: Multi-source decentralized transfer for privacy-preserving BCIs. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 2710–2720 (2022)
4. Wu, W.: Multi-source selection transfer learning with privacy-preserving. *Neural. Process. Lett.* **54**(6), 4921–4950 (2022)
5. Wang, L., et al.: MuKGB-CRS: guarantee privacy and authenticity of cross-domain recommendation via multi-feature knowledge graph integrated blockchain. *Inf. Sci.* **638**, 118915 (2023)
6. Hu, B., et al.: Data protection method against cross-domain inference threat in cyber-physical power grid. *IEEE Trans. Smart Grid* **12**(3), 21–26 (2024)
7. Sahu, A., et al.: Multi-source multi-domain data fusion for cyberattack detection in power systems. *IEEE Access* **9**, 119118–119138 (2021)
8. Shi, M., et al.: Cross-domain privacy-preserving broad network for fault diagnosis of rotating machinery. *Adv. Eng. Inform.* **58**, 102–157 (2023)
9. Li, Y., Hu, X.: Social network analysis of law information privacy protection of cybersecurity based on rough set theory. *Library Hi Tech* **40**(1), 133–151 (2022)
10. Tian, J., et al.: A multi-source information transfer learning method with subdomain adaptation for cross-domain fault diagnosis. *Knowl.-Based Syst.* **243**, 108–110 (2022)
11. Wang, C., et al.: A privacy preservation method for multiple-source unstructured data in online social networks. *Comput. Secur.* **113**, 102–105 (2022)

12. Yang, H., et al.: MTGK: multi-source cross-network node classification via transferable graph knowledge. *Inf. Sci.* **589**, 395–415 (2022)
13. Gong, M., et al.: MISNet: multi-source information-shared EEG emotion recognition network with two-stream structure. *Front. Neurosci.* **18**, 129–132 (2024)
14. He, Q., et al.: Data sharing mechanism and strategy for multi-service integration for smart grid. *Energies* **16**(14), 52–55 (2023)
15. Qiu, G., et al.: Differentiated location privacy protection in mobile communication services: a survey from the semantic perception perspective. *ACM Comput. Surv.* **56**(3), 1–36 (2023)
16. Sun, R., Ren, Y.: A multi-source heterogeneous data fusion method for intelligent systems in the Internet of Things. *Intell. Syst. Appl.* **23**, 200–242 (2024)
17. Sun, P.J.: Security and privacy protection in cloud computing: discussions and challenges. *J. Netw. Comput. Appl. Comput. Appl.* **160**, 47–51 (2020)



Construction of Cold Chain Logistics and Distribution Site Selection System Based on Multi-objective Optimization Model

Hanjie Jia¹, Hua Jiang², and Manjiang Chen¹✉

¹ Kashgar Polytechnic, Kashgar 844000, Xinjiang, China
kszyjsxycmj@126.com

² Binzhou Polytechnic, Binzhou 256603, Shandong, China

Abstract. As the market demand for fresh food and pharmaceutical products grows constantly, the cold chain logistics industry is developing rapidly. An efficient and reasonable distribution network layout is crucial to ensure the quality and safety of merchandise ware, for example, pharmaceutical products. The significance of the cold chain logistics and distribution system is becoming increasingly prominent. A cold chain logistics and distribution site selection system based on a multi-objective optimization model is proposed. A site selection model that integrates multiple factors such as cost, time, environmental impact and service quality is proposed, and the system aims to determine the optimal location of the cold chain logistics facilities by integrating multiple objective functions, such as cost minimization, service coverage maximization, and environmental impact minimization and so on. Adapting Mixed Integer Linear Programming (MILP) to construct the mathematical model and Non-dominated Sorting Genetic Algorithm (NSGA-II) is brought in to solve this complex problem. The system can not only enhance the efficiency of cold chain logistics and distribution with advantage, but also find the optimal or near-optimal solution between multiple conflicting objectives, which can ensure the service quality while effectively reducing the operating costs and carbon emissions, and is incredibly significant for facilitating sustainable development of cold chain logistics.

Keywords: Distribution Site Selection · Cost Optimization · Service Efficiency · Temperature Control Management · GIS Technology · Sustainability

1 Introduction

With the global economy developing and consumers' demand for high quality of life increasing, there is a growing demand for temperature-sensitive commodities such as fresh food and pharmaceutical products. A special form of logistics--cold chain logistics, has proved to be an significant step to make sure of the quality and safety of these commodities, putting forward higher requirements on the location and path planning of distribution centers, and occupying a crucial position in the entire supply chain. In the meantime, cold chain logistics is faced with a flood of twists and turns, for

instance, high cost, strict requirements on temperature control and so on. Therefore, how to scientifically and reasonably plan the cold chain logistics and distribution network, especially the site selection problem has become a hot topic in present study. In order to cope with these challenges, the system is based on the multi-objective optimization theory, which comprehensively considers multiple objectives, such as total logistics cost, low carbon and environmental protection, etc. This study constructs a cold chain logistics distribution site selection system based on a multi-objective optimization model, which achieves efficient, low-cost and reliable operation of the cold chain logistics through selecting site and sketching out path scientifically, and optimization of overall arrangement of distribution centers.

Adopting a variety of advanced algorithms, such as genetic algorithm, mixed integer linear programming (MILP) technology framework combined with genetic algorithm (GA), etc., which makes it possible to quickly find a better solution even in the face of large-scale cases. The system can not only decrease the total cost of logistics and improve the reliability of service, but also meet the requirements of low-carbon green and environmental protection, providing strong support for the sustainable development of enterprises.

The construction of cold chain logistics distribution site selection system based on multi-objective optimization model is a efficacious avenue to cope with the challenges of cold chain logistics, which provides an effective tool for the related enterprises, which can realize the efficient, low-cost and reliable working of cold chain logistics, better cope with the competitive pressure in the market, and provide customers with better distribution services, and at the same time, it also provides a theoretical support and technical methods to develop cold chain logistics industry towards a more efficient and green direction. It also provides theoretical support and technical means to push the development of cold chain logistics industry in a more efficient and green direction.

2 Related Works

In recent years, academic research on cold chain logistics and distribution siting system has been in-depth, and in the domain of cold chain logistics, optimized siting and path planning is the key to improve efficiency and reduce costs. Researchers have proposed a method combining greedy algorithm and immune optimization strategy to work out the issue of cold chain logistics facility siting and path strategy with time window restriction. By introducing the concepts of clone selection and antibody diversity in immune system and integrating the characteristics of greedy algorithm to locate the local optimal solution quickly, the method aims to find the logistics network layout plan that meets the requirements of time window of cold chain transportation and has the lowest cost, which can effectively tackle the mishap and has good application prospects [1]. A data collection intelligent warehouse monitoring system based on Internet of Things (IoT) technology is investigated, which utilizes sensors to monitor environmental parameters (such as temperature and humidity) and cargo status in the warehouse in real time, and improves the efficiency of warehouse management and guarantees the safety of goods through data analysis. In addition, how to use the cloud computing platform to process large-scale data sets and optimization suggestions for different application scenarios are also discussed to further enhance the performance and practicality of the system [2]. Researchers

have focused on the training mode of logistics professionals in the context of vocational education, emphasizing the importance of combining theoretical learning and practical operation, and proposing a series of specific implementation strategies, including the school-enterprise cooperation mechanism, which is designed to promote the enhancement of students' employability while helping enterprises to obtain the professional skills they need [3]. There is a study on the cold chain logistics location-path problem under the "to the counter" mode taking the carbon emission element into account, and a mathematical model that integrates the dual objectives of cost-effectiveness and environmental protection is developed and solved by mixed-integer linear programming method, which ensures that while ensuring the quality of service, the new model can effectively diminish the carbon emission of overall operation process and thus achieve a more environmentally friendly and efficient cold chain logistics. The new pattern can effectively reduce the carbon emissions during the overall operation process to realize a more environmentally friendly and efficient cold chain logistics service [4]. Researchers focused on the location of cold chain logistics and distribution centers for fresh agricultural products and their subsequent path planning, established a model with cost minimization as the Goal function, and used genetic algorithms to solve this complex optimization problem, the proposed model and algorithms can effectively reduce the operating costs under the premise of ensuring product freshness [5]. The impact of node disruption and carbon emission on cold chain logistics network planning is explored, and the study adopts stochastic planning method to deal with the uncertainty factors, and takes the carbon footprint as one of the additional constraints, and the appropriate increase of redundant paths can significantly improve the network's risk-resistant ability and also achieve a certain degree of emission reduction by optimizing the selection of paths, which is of great significance for the reconstruction of a more robust and low-carbon cold chain logistics network structure. It is of great significance to construct a more robust and low-carbon cold chain logistics network structure [6]. In the field of cold chain logistics, numerous research papers have probed the optimization of distribution center locations, route planning, and real-time monitoring technologies to enhance efficiency and reduce carbon emissions. Researchers have proposed a low-carbon emission competitive urban cold chain logistics distribution center location optimization method, aiming to determine the optimal layout of logistics nodes by comprehensively considering cost and environmental impact [7]. Some studies have focused on improving the energy efficiency of perishable food supply chains through the use of real-time temperature and humidity monitoring technology, thereby reducing energy waste [8]. Scholars have analyzed how to maintain the stability of perishable goods in cold logistics chains, pointing out that effective temperature control measures are crucial for maintaining product quality [9]. A hybrid algorithm has been developed to address the issue of cold chain logistics distribution center location, combining all kinds of intelligent optimization strategies to find the best solution [10]. Further research has delved into the optimization of fresh agricultural product distribution paths and resource scheduling under low-carbon environmental constraints [11]. Multi-objective optimization studies have been conducted on green hub locations for multi-temperature joint distribution of perishables, with the aim of simultaneously lowering operational costs and environmental footprints [12]. A multi-objective optimization pattern for the selection of cold chain logistics distribution center

locations based on carbon emissions has been proposed, emphasizing the importance of environmental protection [13], and methodologies for designing cold chain logistics networks involving heterogeneous fleets under carbon reduction policies have been discussed [14]. Efforts have been made to design a matching mechanism for partners in Hainan’s cold chain logistics using multi-objective optimization methods to promote efficient collaboration within the region [15]. An optimal path selection algorithm suitable for rural multi-temperature zone cold chain logistics has been put forward, aimed at balancing transportation costs and time efficiency among other factors [16].

These studies demonstrate a incredibly extensive range of points of interest in the field of cold chain logistics, including, but not limited to, site optimization, route planning, application of real-time monitoring technologies, and a focus on environmentally friendly solutions. Through these efforts, it is expected that cold chain logistics will become more efficient and sustainable in the future.

3 Methods

3.1 System Architecture

The cold chain logistics and distribution site selection system based on multi-objective optimization model needs to consider multiple factors and provide an optimal solution. The overall architecture used in this system is shown in Fig. 1:

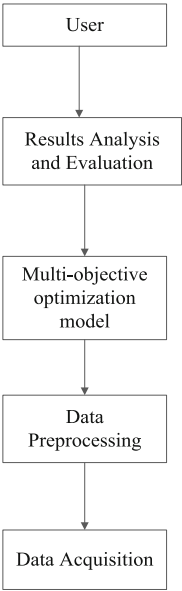


Fig. 1. Overall system architecture

The overall architecture of the system is shown in Fig. 1, which allows users to easily input parameters, view results and interact with each other, supports map visualization, displays the locations of potential distribution centers and their service areas, and provides report generation functions to output detailed information and suggestions for optimization schemes. The input interface allows the user to enter necessary parameters such as demand distribution, cost data, etc. The visualization interface uses the Map API to display the location of potential distribution centers and their service areas. The visualization interface uses a map API to display the distribution center locations and service areas. Optimization results are displayed in the form of charts, tables, etc., including key indicators such as total cost, service level, environmental impact, etc. Decision makers can view and compare the results of different site selection options. Operators can enter data and parameters, start the optimization process, perform multi-dimensional analysis of optimization results, including economy, efficiency and service quality, etc., and perform sensitive analysis to appraise the influence of changes in key parameters on the final Sensitivity analysis is used to estimate the impact of changes in critical parameters on the final solution, allowing users to compare the advantages and disadvantages of different solutions. The analyst is responsible for analyzing the optimization results in depth to provide a basis for decision making, and the decision maker makes the final decision based on the fructify of the analysis. Researchers are responsible for constructing and validating optimization models, and technologists implement and tune optimization algorithms. Cleans and formats data, removes noise and inconsistencies, integrates data from different sources, ensures data consistency and integrity. Data collection is to obtain historical order data, inventory data, etc. from internal systems (e.g., ERP, WMS), collect real-time logistics data through IoT devices (e.g., temperature sensors, GPS trackers), and obtain external data, such as traffic conditions, weather forecasts, and market demand forecasts. Connecting to internal information systems to extract relevant data, accessing external data through APIs or data subscription services, and collecting real-time data using IoT devices. Data collection engineers are responsible for configuring and maintaining the data collection system, and IT support staff make sure of the average working of the data collection system.

The system architecture diagram shows the various levels and functions of the cold chain logistics distribution site selection system, from data collection to data preprocessing, to the construction of multi-objective optimization model and result analysis, which makes the system more modular and easy to develop, maintain and expand. Cold chain logistics enterprises can make more scientific and efficient decisions on distribution center location and improve the overall operational efficiency and service level.

3.2 Model Training and Optimization

The construction of the cold chain logistics and distribution site selection system based on the multi-objective optimization model involves several steps such as data preparation, model selection. Data preparation and preprocessing is to assemble record related to cold chain logistics and distribution site selection. Obtain raw data from multiple data sources, process missing values, outliers, and duplicates. Select a model framework suitable for multi-objective optimization problems, define the optimization objective function, decision variables and constraints, and design the model structure to ensure

that it can effectively solve practical problems. Define the optimization objectives, determine the decision variables that need to be optimized, define the constraints that must be complied with, initialize the parameters in the pattern to provide the initial values for the subsequent training process, set the hyper-parameters of the algorithm, set the initial parameters based on experience and preliminary experimental results, and select appropriate hyperparameter values in order to obtain better performance in the training process. Select the most suitable algorithm according to the characteristics of the issue, the algorithm implementation is to write and debug the algorithm code to ensure its correctness and effectiveness, use the selected algorithm and initialized parameters to train

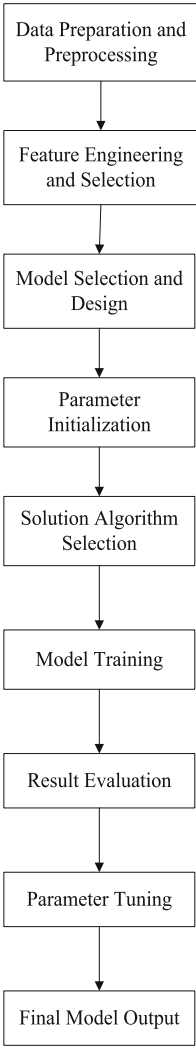


Fig. 2. Flowchart of model training and optimization

the model, and in the training process, iteratively, optimize the model parameters step by step to find the optimal solution, and gradually approach optimal solution by updating the model parameters through multiple iterations. Set the convergence condition and stop training when the condition is satisfied. Calculate key performance indicators. Use charts and reports to display the evaluation results to help users understand performance of the model, and adjust the model parameters according to the evaluation results to further optimize the model performance. Balance the model parameters based on the evaluation results, and adjust the hyperparameters of the algorithm to improve the solving effect. Output the final trained model for cold chain logistics and distribution site selection decision, model documentation and usage guidelines to help users understand and apply the model.

The model training and optimization flowchart shows the whole process from data preparation to final model output, with specific tasks and responsibilities for each step, ensuring the efficiency and scalability of the model training and optimization process, and constructing a powerful multi-objective optimization model to improve the overall operational efficiency and service level (Fig. 2).

4 Results and Discussion

4.1 Results

For the sake of demonstrating the results of the cold chain logistics and distribution site selection system based on a multi-objective optimization model, the table compares key metrics under different site selection scenarios, including the total cost, the service level, the environmental impact, and whether or not each potential location is selected as a distribution center. Table 1 shows the specific metrics recorded in this experiment and the results observed.

Table 1. Specific indicators and experimental results

Location Plan	Logistics Cost (Ten Thousand Yuan)	Service Reliability Score (out of 100)	Carbon Emissions (tons/year)
Plan A	1200	85	300
Plan B	1150	90	320
Plan C	1250	95	280
Plan D	1180	88	295
Plan E	1220	92	270

Different site options are listed, each of which indicates a possibility of establishing a cold chain logistics and distribution center at a specific location. The logistics cost (RMB 10,000) indicates the total operating cost of the cold chain logistics and distribution network under each location option. The lower the value means the lower the cost, the more advantageous the option is in terms of cost, e.g., the logistics cost of Option

B is RMB 11,500,000, which is one of the most cost-effective options. The service reliability score is based on the reliability of the distribution process, and demonstrate the performance of each site selection option in terms of service reliability by means of scoring, which is based on a flood of elements. The higher the score, the better the service just for reliability. For example, Option C has a service reliability score of 95, making it the most reliable option in terms of service. Carbon Emission (tons/year) lists the total carbon emission generated by each location option in the procedure of cold chain logistics and distribution, the lower the value means the lower the impact on the environment, the better the performance in environmental protection, for example, the carbon emission of Option E is 270 tons/year, which is the option with the smallest environmental impact.

Table 2. The coefficients and significance test results of the negative binomial regression model

Variable Name	Coefficient (β)	Standard Error (SE)	Z-value	P-value	Significance Level
Intercept	0.50	0.12	4.17	<0.001	***
Total Cost	-0.03	0.01	-2.89	0.004	**
Service Level	0.10	0.03	3.33	0.001	**
Environmental Impact Index	-0.05	0.02	-2.50	0.012	*
Location A	0.60	0.20	3.00	0.003	**
Location B	-0.40	0.18	-2.22	0.026	*
Location C	0.30	0.15	2.00	0.045	*

From the perspective of logistics cost, if the decision maker mainly focuses on cost, then Scenario B (11.5 million yuan) is the first choice because it has the lowest logistics cost; if the decision maker values service reliability more, then Scenario C (95 points) is the best choice because it has the highest service reliability score; if the decision maker wants to balance cost, service reliability and environmental impact, it needs to comprehensively consider these three factors. For example, Scenario E achieves a relatively good balance between logistics costs (\$12.2 million), service reliability (92 points) and carbon emissions (270 tons/year) and is a compromise. Decision-makers also need to consider other non-quantifiable factors, such as policy environment, social impacts, geographical accessibility, etc., which are difficult to quantify but equally important.

In the cold chain logistics and distribution site selection system, a negative binomial regression model is used to analyze the relationship between certain factors (e.g., cost, service level, environmental impacts, etc.) and the site selection decision, and a table is constructed to show the coefficients of the regression model and the results of their significance tests. The coefficients of the negative binomial regression model. The significance test results are displayed in Table 2.

This experiment is the independent variables in the regression model and their corresponding regression coefficients, standard errors, z-values, p-values and significance. The following are the specific analyses for each of the independent variables:

The variable names list all the independent and dependent variables used in the regression analysis. The coefficients (β) indicate the extent to which each independent variable affects the dependent variable, with positive numbers indicating a positive effect and negative numbers indicating a negative effect. Standard Error (SE) is the standard error of the estimated coefficient and is used to measure the precision of the estimated coefficient. z-value is the statistic obtained by dividing the coefficient by its standard error and is used for hypothesis testing. p-value is the probability value calculated from the z-value and is used to determine whether the coefficient is statistically significant. Significance levels are marked with an asterisk (*) to indicate different levels of significance: *** means $p < 0.001$, ** means $p < 0.01$, * means $p < 0.05$, and the absence of an asterisk indicates non-significance ($p > 0.05$).

When all independent variables are zero, the expected value of the dependent variable is 0.50 and this value is statistically significant ($p < 0.001$). Whenever the total cost increases by one unit, the dependent variable decreases by 0.03 units accordingly and this negative correlation also reaches a statistically significant level ($p = 0.004$). The level of service increased by 0.10 units for each percentage point increase in the dependent variable and again this was statistically significant ($p = 0.001$). For each unit increase in ESI the dependent variable decreases by 0.05 units, which is statistically significant ($p = 0.012$). Position A, B, C, D, E These variables represent the selection of different positions, each with its own specific coefficient of influence, for example, the selection of position A would result in an increase of 0.60 units in the dependent variable, whereas the selection of position B would result in a decrease of 0.40 units in the dependent variable.

The coefficient of the intercept is 0.50, the standard error is 0.12, the Z-value is 4.17, the p-value is less than 0.001, and the significance level p is less than 0.001, the expected value of the dependent variable when all independent variables are zero is 0.50, and this intercept is statistically very significant, implying that a base probability exists even if there is no effect of other factors. The coefficient of total cost is -0.03 with a standard error of 0.01, a z-value of -2.89 , a p-value of 0.004 and a significance level of p less than 0.01, for every unit increase in the total cost the dependent variable decreases by 0.03 units, which indicates that higher costs reduce the likelihood of site selection, and a p-value of less than 0.05 indicates that this relationship is statistically significant. The coefficient of level of service is 0.10, standard error is 0.03, z-value is 3.33, p-value is 0.001, and significance level is p less than 0.01, for every percentage point increase in the level of service the dependent variable increases by 0.10 units, which means that higher level of service increases the likelihood of site selection, and the p-value is less than 0.01 shows that the relationship is highly statistically significant. The coefficient of the ESI is -0.05 and its standard error is 0.02, which corresponds to a z-value of -2.50 and a p-value of 0.012. With a significance level of p less than 0.05. Each unit increase in the environmental impact index decreases the dependent variable by 0.05 units, which means that greater environmental impacts decrease the likelihood of site selection, and a p-value of less than 0.05 indicates that the relationship is statistically

significant. Location A has a coefficient of 0.60, a standard error of 0.20, a z-value of 3.00, a p-value of 0.003, and a significance level of p less than 0.01, choosing location A as the distribution center dependent variable increases by 0.60 units indicating that location A has a positive impact, and a p-value of less than 0.01 shows significance. The coefficient of location B is -0.40 , standard error is 0.18, z-value is -2.22 , p-value is 0.026 and significance level is p less than 0.05, choosing location B decreases the dependent variable by 0.40 units indicating that location B has a negative impact on site selection, p-value is less than 0.05 but greater than 0.01 showing moderate significance. Location C has a coefficient of 0.30, standard error of 0.15, z-value of 2.00, p-value of 0.045, and a significance level of p less than 0.05, choosing location C increases the dependent variable by 0.30 units indicating that location C has a positive impact, p-value is slightly less than 0.05 showing marginal significance.

The coefficients of each variable reflect the direction and magnitude of the influence on the location decision, and significance tests determine whether these influences are reliable.

4.2 Discussion

This study concludes the trade-off between cost and service level through the analysis of different site selection options. Choosing a location as a distribution center that can effectively reduce carbon emissions not only meets the requirements of the increasingly stringent environmental protection regulations but also can also enhance corporate image; the choice of location has a decisive role in the efficiency of the entire cold chain logistics system, especially in terms of shortening transportation distances and reducing losses.

This study enriches the multi-objective optimization theory in the field of cold chain logistics, provides a new perspective to understand and solve the complex logistics decision-making problems, and the research results are directly applied to the actual cold chain logistics network planning, which helps enterprises to make site selection decisions more scientifically so as to realize the maximization of economic and social benefits.

The cold chain logistics distribution site selection system based on the multi-objective optimization model not only provides an effective tool for the current cold chain logistics management, but also points out the direction for future research in this field. With the advancement of technology and the development of methodology, more innovative research results are expected to emerge in this field.

5 Conclusion

The reconstruction of a cold chain logistics and distribution site selection system is a complex and critical process, which is in direct relation to the efficiency and cost of cold chain logistics. Through the review of existing literature and the analysis of practical requirements, an optimization framework capable of handling multiple conflicting objectives is designed and solved with appropriate algorithms.

Through analyzing a case of a cold chain logistics and distribution network in a specific region, the effectiveness of the proposed model is displayed. The results show that the new model is capable of significantly improving the performance of the whole system compared to the traditional approach. The optimized cold chain logistics and distribution network not only reduces the operation cost, but also improves the service efficiency, which helps enterprises to enhance market competitiveness. By reducing carbon emissions and other environmental pollution, the new model meets the requirements of sustainable development and is conducive to environmental protection. Despite some important results of this study, there are still some restrictions. Validity of the model is highly dependent on the quality and accuracy of input data, which are often difficult to obtain or not completely reliable. Uncertainties in reality, such as weather changes and traffic conditions, may affect the usefulness of the model. Although genetic algorithms are capable of handling multi-objective optimization problems, they still face high computational complexity when dealing with large-scale instances, which limits their wide application.

The reconstruction of cold chain logistics and distribution site selection system based on multi-objective optimization model reduced operating costs. Efficient cold chain logistics and distribution services can boost customer satisfaction, enhance the competitiveness of enterprises, and promote the development of cold chain logistics. Construction and improvement of the system can help to promote the standardization of the cold chain logistics industry and improve the service level and efficiency of the whole industry.

With continuous growth of related technologies and theories, the future cold chain logistics network will become more efficient, green and adaptable.

Acknowledgments. This article is the research outcome of the science and technology project “Research and Demonstration of Key Technologies for Fresh Agricultural Products Preservation and Transportation” in the Kashgar region of Xinjiang (Project Number: KS2023019), led by Chen Manjiang.

References

1. Ning Y-J., Zhang, H-Z.: Greedy immune optimization algorithm for solving cold chain logistics site selection path problem with time window. *Logistics Science and Technology*, **47**(08), 140–146 (2024). <https://doi.org/10.13714/j.cnki.1002-3100.2024.08.036>
2. D Jiuling Z Wanmeng C Gang 2023 Intelligent warehouse monitoring system based on data acquisition *Internet Things Technol.* 13 12 143 146 <https://doi.org/10.16667/j.issn.2095-1302.2023.12.038>
3. Yang, Z.P., Li, F.M., Zhang, L.: Research on vocational talent cultivation mode of logistics specialty oriented by industry-teaching integration. *J. Hebei Softw. Vocat. Tech. Coll.* **25**(04), 37–40+46 (2023). <https://doi.org/10.13314/j.cnki.jhbsi.2023.04.015>
4. Liu, Y., Wang, N., Chen, C.: Location-path problem of “to-the-counter” mode cold chain logistics considering carbon emission. *Logistics Sci. Technol.* **46**(07), 130–134+148 (2023). <https://doi.org/10.13714/j.cnki.1002-3100.2023.07.031>
5. FU Xiaoting TANG Qiusheng 2023 Site selection-path planning for cold chain logistics and distribution center of fresh agricultural products *Transp. Sci. Econ.* 25 04 11 19 <https://doi.org/10.19348/j.cnki.issn1008-5696.2023.04.002>

6. MG Zeng XH Wang JY Lai 2021 A study on cold chain logistics network planning considering node disruption and carbon emission J. South China Univ. Technol. (Soc. Sci. Ed.) 23 04 55 66 <https://doi.org/10.19366/j.cnki.1009-055X.2021.04.006>
7. Zhang, S.Y., Chen, N., She, N., Li, K.: Low-carbon emission competitive distribution center location optimization for urban cold chain logistics. Comput. Ind. Eng. **154** (2021). <https://doi.org/10.1016/j.cie.2021.107120>
8. Aguiar, M.L., Gaspar, P.D., Silva, P.D., Domingues, L.C., Silva, D.M.: Real-time monitoring of temperature and humidity for perishable food distribution to enhance supply-chain energy efficiency. Processes, **10**(11) (2022). <https://doi.org/10.3390/pr10112286>
9. M Bogataj L Bogataj R Vodopivec 2005 Maintaining the stability of perishable goods in cold logistics chains Int. J. Prod. Econ. 93–4 345 356 <https://doi.org/10.1016/j.ijpe.2004.06.032>
10. SH Dou GY Liu YB Yang 2020 A Novel hybrid algorithm for solving the cold chain logistics distribution center location problem IEEE Access 8 88769 88776 <https://doi.org/10.1109/access.2020.2990988>
11. Fu, Q.M., Li, J., Chen, H.H.: Optimizing fresh agricultural product distribution paths with resource scheduling under low-carbon environmental constraints. Sci. Program. **2022** (2022). <https://doi.org/10.1155/2022/7692135>
12. M Golestani SH Moosavirad Y Asadi S Biglari 2021 Multi-objective optimization of green hub locations for multi-temperature joint distribution of perishables in a cold supply chain Sustainable Prod. Consumption 27 1183 1194 <https://doi.org/10.1016/j.spc.2021.02.026>
13. XG Li K Zhou 2021 Carbon emission-based multi-objective optimization for cold chain logistics distribution center location Environ. Sci. Pollut. Res. 28 25 32396 32404 <https://doi.org/10.1007/s11356-021-12992-w>
14. Liu, Y.H., Shi, X.L.: Designing a cold chain distribution network with a heterogeneous fleet under carbon reduction policies. Transp. Plan. Technol. (2024). <https://doi.org/10.1080/03081060.2024.2375295>
15. Mo, H.P., Deng, C., Chen, Y.T., Huang, Y.C.: Matching mechanism for partners in Hainan's cold chain logistics using multiobjective optimization. Math. Probl. Eng. **2022** (2022). <https://doi.org/10.1155/2022/5506338>
16. Qi, C.W.: An algorithm for optimal path selection in rural multi-temperature zone cold chain logistics using multi-objective optimization. Int. J. Comput. Intell. Syst. **17**(1) (2024). <https://doi.org/10.1007/s44196-024-00616-3>



Hole Detection Algorithm Based on Channel Fusion Siamese Network

Nuan Sun¹(✉), Chunhe Shi², Yanchao Cui¹, Yaran Wang¹, Xiaoying Shen¹,
and Xinru Shao¹

¹ Department of Artificial Intelligence, Shenyang University, Shenyang, China
sn97514325@163.com

² Institute of Innovation Science and Technology, Shenyang University, Shenyang, China

Abstract. With the increasing popularity of shooting sports, traditional manual readings have defects such as missed reports, false reports, and low security. To solve the existing problems, this paper utilizes deep learning technology to propose a dual-channel fusion Siamese network and detection algorithm for solving the difficult problems of microscopic hole position recognition and ring numbers reading on the target surfaces. By utilizing adaptive feature learning, combined with the closed area algorithm and ring value matrix, the calibration error in traditional methods is solved, achieving precise hole positioning and accurate ring number reading. The experimental findings indicate the high efficiency and accuracy of the above methods in target image processing and hole detection.

Keywords: Hole Detection · Channel Fusion · Siamese Networks · Ring Area Calibration

1 Introduction

In recent years, shooting sports have increasingly evolved into a widely recognized competitive activity, accompanied by the proliferation of shooting clubs, which indicates significant market potential. A critical aspect of shooting sports is the precise assessment of hole rings; however, traditional manual counting methods pose safety risks, as well as issues such as missed counts, inaccurate reports, and inefficiencies. Among the automatic reporting equipment currently available, most rely on hardware-based designs that entail high costs and specific environmental requirements.

This study employs deep learning technology to resolve the fundamental challenges with the recognition of micro hole positions and the reading of scoring rings. By identifying the limitations of existing methodologies, we propose effective optimization theories and techniques. Furthermore, we enhance and refine the prevailing algorithmic models to increase the accuracy of hole detection on targets.

This technology represents a crucial component of automatic reporting devices, offering a cost-effective and reusable solution that can adapt to various target surfaces, environmental conditions, and shooting specifications. Consequently, it addresses a significant gap in the industry and has potential applications across multiple domains, including military and sports contexts.

2 Related Works

In recent years, the image processing and recognition techniques have been developed rapidly and these technologies are already being applied to identify the numbers of holes on the target surface. This research describes the image processing method based on MATLAB analysis tool. Compared with traditional methods, MATLAB provides more advanced image processing functions, including image cropping and resizing, noise reduction, blur removal and image enhancement [1].

This thesis adopted the method of deep learning to study the hole identification and the assessment of ring value. Siamese Neural Networks have been applied to capture the similarity of items within the image domain. This study reviews the intersection of these two fields, namely how Siamese Neural Networks are used for recommendations [2, 3].

The Converter, a neural network utilizing a self-attention mechanism, is significantly superior to the traditional convolutional and recursive models in a range of visual processing applications. Therefore, more efficient and direct solutions for many segmentation tasks are provided by the Vision Converters [4].

Deep Neural Network (DNN) models are rapidly evolving on both resource-rich settings and resource-limited devices [5]. Compared with other popular deep learning architectures, Residual Network (ResNet) has shown remarkable effectiveness in evaluating various performance metrics, including sensitivity, specificity, accuracy, and F1 scores [6].

Sobel edge detection is a popular technique in computer vision and image processing. This review introduces a completely software-based approach that offers a more straightforward and cost-effective alternative. This algorithm minimizes the number of arithmetic operations and data load, thereby improving the processing speed and reducing energy consumption [7].

In this paper, an enhanced adaptive superpixel segmentation technique using a regression prediction model is proposed. This method solves the problem that the segmentation edge of the traditional superpixel method cannot fully fit the defect target edge [8].

3 Methods

Considering the specific characteristics of the target surface, a hole detection algorithm based on channel fusion Siamese network is proposed to accurately obtain the location of the holes. At the same time, the annular value region of the target surface was calibrated to generate an annular matrix. The human-computer interaction interface was designed by combining the hole position with the annular matrix to facilitate a more intuitive interpretation of the ring count.

3.1 Dataset Collection

Since there is no publicly available data set about the target surface, this paper uses Python code to intercept all the obtained high-resolution video of the target surface and the corresponding 2, 4 and 8 times downsampled video according to a fixed number of frames, so as to produce data sets with different magnifications. In order to expand the data set and ensure the size of the data set, the obtained samples were flipped and rotated.

3.2 Hole Detection Algorithm Based on Channel Fusion Siamese Network

- Channel fusion Siamese neural network

To more effectively determine the area of the hole on the target surface following the hit, this study introduces a channel fusion Siamese network, the architecture of which is illustrated in Fig. 1.

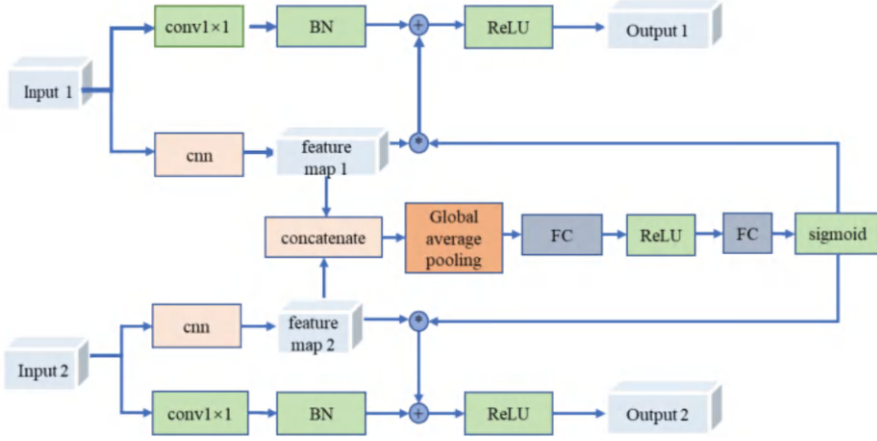


Fig. 1. Structure of channel fusion Siamese network

The network employs a Siamese network [2, 3] as its foundational framework, wherein input image blocks derived target images from the hit are processed by two branches that share weights. To establish a connection between the channel features of these two branches, this study employs a residual structure as the feature extraction network for both branches of the Siamese network. The feature maps produced by the distinct branches are connected along the channel latitude. Subsequently, a channel attention mechanism is applied to identify and prioritize the channels that enhance detection outcomes, allowing for the reallocation of weights across the feature map channels.

The channel fusion Siamese network model primarily comprises residual structures [9], as shown in Fig. 2a), along with channel fusion modules, depicted in Fig. 2b).

As illustrated in Fig. 2a), let the inputs of the two residual blocks be $I_1 \in \mathbb{R}^{h \times w \times c'}$ and $I_2 \in \mathbb{R}^{h \times w \times c'}$, and let $F(\cdot, W)$ be the mapping through the convolutional network that converts the input information into feature maps $F_1 \in \mathbb{R}^{h \times w \times c}$ and $F_2 \in \mathbb{R}^{h \times w \times c}$ with channel numbers c and c' . If they are simultaneous, $h(\cdot)$ is the identity mapping; if they are not simultaneous, $h(\cdot)$ can be scaled by channel number through 1×1 convolution. The input feature map $I \in \mathbb{R}^{h \times w \times c'}$ is transformed into $\tilde{I} \in \mathbb{R}^{h \times w \times c}$ by the mapping $h(I)$. In Fig. 2b), and feature map V_1 with the same size and channel number $2c$ is obtained by fusing $F_1 \in \mathbb{R}^{h \times w \times c}$ and $F_2 \in \mathbb{R}^{h \times w \times c}$ in the channel dimension.

As illustrated in Fig. 2b), the channel fusion module initially integrates the two input feature maps F_1 and F_2 along the channel dimension to obtain feature map V_1 . Subsequently, a global average pooling operation is applied to transform these feature maps

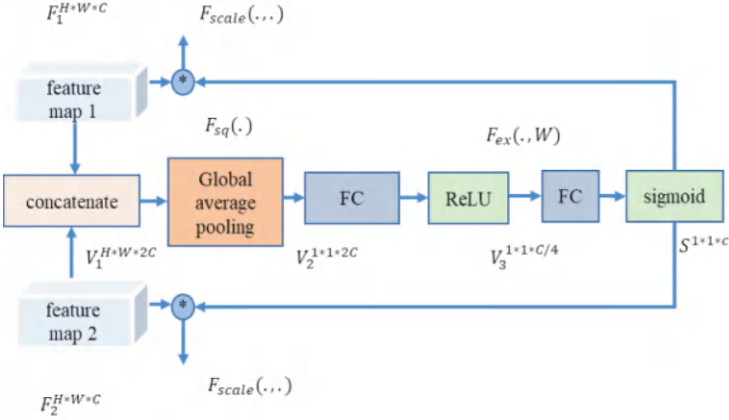
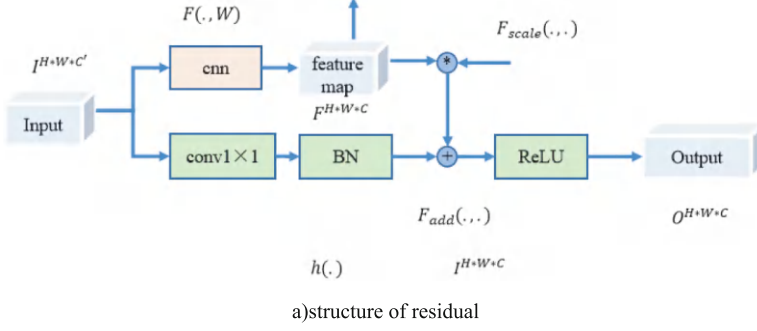


Fig. 2. Residual structure and channel fusion module

into $1 \times 1 \times c$ feature vector V_2 . This is followed by the establishment of inter-channel connections through two 1×1 convolutional layers, which facilitate the reassignment of channel weights. Following two convolutional operations, the quantity of channels within the feature maps become $c/4$ and c respectively. The channel attention mechanism [4] then adaptively adjusts the weight of each channel by creating connections between the feature maps F_1 and F_2 and all channels, thereby enabling the learning of relevant inter-channel information.

- Hole detection network based on channel fusion Siamese network

On network, which is predicated on the channel fusion Siamese network, is illustrated in Fig. 3. This study investigates a deep learning approach utilizing ResNet32 [6] for the purpose of hole detection. ResNet32 is characterized by its 32 layers dedicated to parameter learning, devoid of pooling layers, with the potential for size reduction achieved through modifications to the stride of the convolution operation. Consider bullet hole detection as a task involving pixel-wise classification, necessitating the processing of the feature map at the channel level. The network architecture is comprised of three segments, each formed by stacked residual modules, which regulate downsampling by

modifying the stride of the convolution operation in the initial layer. Furthermore, the channel-fusion Siamese network framework is incorporated, wherein the feature map undergoes downsampling through adjustments to the stride of the convolution operation within the fusion module of a designated layer.

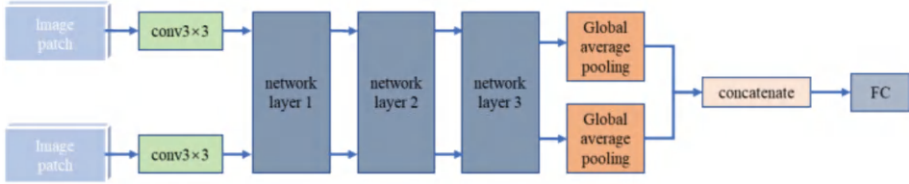


Fig. 3. Structure of change detection algorithm based on channel fusion Siamese network

Following the hit incident, the captured images are transmitted to a dual neural network framework designed for channel fusion in hole detection. The output generated by this network, denoted as (d_i) , possesses dimensions of $h \times w \times 2$. The probability of a pixel point in the image remaining unchanged is $d_1^{h \times w}$, while the probability of a pixel point undergoing a change is $d_2^{h \times w}$.

Ring area calibration

To determine the quantity of rings present in the region corresponding to the location of the hole. It is essential to not only consider the data pertaining to the position of the hole but also to acquire the range corresponding to each ring value.

1) Loop edge detection

Converting the chest ring target image into a grayscale image as the input for edge detection, this study employs the Sobel operator to obtain edge information from the target image. Typically, the edge and orientation of a pixel located at coordinates (x, y) within an image represented by f are characterized by the gradient Δf , as shown in Eq. (1).

$$\nabla f = \text{grad}(f) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}. \quad (1)$$

The formulas for the amplitude and direction of the gradient vector are presented in Eqs. (2) and (3).

$$g(x) = \sqrt{g_x^2 + g_y^2} \approx |g_x| + |g_y| \quad (2)$$

$$\sigma(x, y) = \arctan \begin{bmatrix} g_x \\ g_y \end{bmatrix} \quad (3)$$

The Sobel operator [7] performs edge extraction in two directions, vertical and horizontal. This procedure is shown in Fig. 4.

To compute the approximate derivatives in both vertical and horizontal directions, as shown in Eqs. (4) and (5), one should perform a multiplication of the image feature

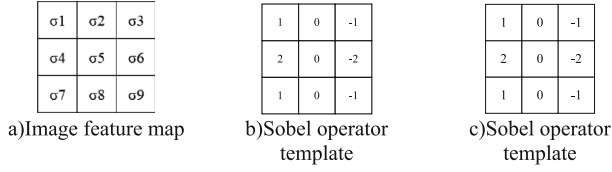


Fig. 4. 3×3 image and Sobel operator template

map depicted in Fig. 4a) with the Sobel operators located at the corresponding positions in Figs. 4b) and Fig. 4c). Subsequently, the results of these multiplications should be summed.

$$g_x = (\sigma_1 + 2\sigma_4 + \sigma_7) - (\sigma_3 + 2\sigma_6 + \sigma_9) \quad (4)$$

$$g_y = (\sigma_1 + 2\sigma_4 + \sigma_7) - (\sigma_3 + 2\sigma_6 + \sigma_9) \quad (5)$$

2) Morphological processing

Following the edge detection process, the resulting image contains interfering edge information, including artifacts such as holes and double loops. Consequently, additional processing is required to eliminate extraneous edge information and bridge the gaps between the double ring lines. Initially, the image undergoes a dilation operation, which effectively reduces the size of the irrelevant edge information to small areas while simultaneously transforming the double ring lines into a more pronounced thick ring line. Given that the small areas produced by the dilation are significantly smaller than the area occupied by the ring line region, this characteristic can be utilized to remove the small areas, thereby preserving the integrity of the complete ring line.

3) Loop midline extraction

The target line derived in step 2) exhibits a considerable thickness, and directly marking the ring area may result in significant inaccuracies. To address this issue, the present study employs the morphological processing function [1] `[bwnorp h(BW, operation, n)]` in MATLAB to extract the midline of the loop.

4) Ring area marker

Following the extraction of the central line of the ring, various independent closed ring areas were delineated, these distinct ring areas were identified and assigned different colors. This study presents a methodology for marking ring areas [8] that is predicated on the area of each ring. Initially, the area of each closed ring is computed, and the results are organized in ascending order from largest to smallest. Subsequently, utilizing the area characteristics of the ring regions on the target surface, each closed region is designated with actual ring numbers, resulting in the formation of a ring area marking matrix.

Following the aforementioned series of operations, each annular region of the target surface has been delineated, and the corresponding ring values have been designated. By integrating this information with the coordinates of the centroid of the hole, it becomes

possible to ascertain the ring number associated with the hole. In practical target hitting scenarios, a particular circumstance may arise in which the hole is precisely aligned with the target line. In such cases, this study adopts the convention of considering the side with the higher ring value as the valid score.

4 Result and Discussion

4.1 Data Analysis

This study will perform experiments utilizing the channel attention fusion Siamese network detection algorithm alongside DNN [5] and ResNet32, all under identical conditions. A comparative analysis of the results will be presented, with quantitative metrics detailed in Table 1.

Table 1. Different Methods on the Data set Numerical Index

	Precision	Recall	F1	Accuracy
DNN	0.902	0.897	0.899	0.969
ResNet	0.931	0.913	0.922	0.978
Textual algorithm	0.950	0.954	0.952	0.989

The findings of the experiment indicate that the hole detection algorithm introduced in this study attains a recognition rate of 0.989 for the identification of hole positions.

4.2 Ring Test Experiment

In order to minimize computational complexity, the processed grayscale [10] target image is utilized as the input for edge detection. The input image is shown in Fig. 5a). The Sobel operator is employed to extract edge features from this input image, with the resulting output presented in Fig. 5b).

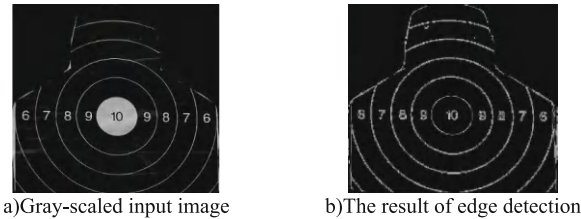


Fig. 5. Edge detection comparison chart

Following the application of the Sobel operator for edge extraction, the resultant image retains the full outline of the humanoid Fig as well as the loop of the chest ring

target. However, it also incorporates extraneous elements such as holes and character information. Notably, the loop primarily consists of both the inner and outer lines, necessitating the removal of these interfering details, particularly the textual elements. To initiate this process, the image undergoes an inflation procedure, with the outcome shown in Fig. 6.

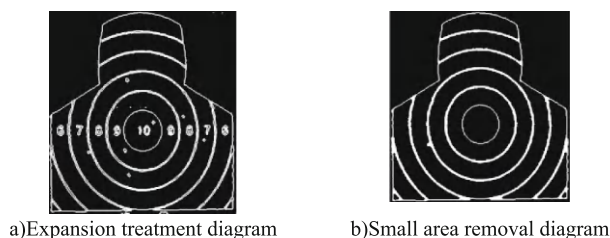


Fig. 6. The inflated image

In the Fig, the enlarged characters and holes delineate distinct small regions that possess a significantly smaller volume compared to the annular regions. This characteristic allows for the elimination of these minor areas while preserving the integrity of the entire ring; however, the resultant ring exhibits a markedly coarser texture.

This study employs the morphological processing functions available in MATLAB, specifically the command `bwmorph h(BW, operation, n)`, to facilitate the extraction of the ring skeleton. The outcomes are shown in Fig. 7.



Fig. 7. The effect of the middle line treatment

The Fig shows that, following processing, the ring has undergone a reduction in thickness while preserving a significant level of structural integrity. Additionally, the ring area has been delineated into distinct closed regions, which are identified and color-coded. This outcome is presented in Fig. 8a).

In Fig. 8a), the ring values corresponding to the head and the left/right lower corners of the thoracic target are identical, yet they are represented in different colors. The ring values for the chest ring target are sequentially numbered from 5 to 10, progressing from the outermost to the innermost ring, with the outermost ring designated as 5. The regions identified as non-humanoid and the left/right lower corner areas are classified as invalid

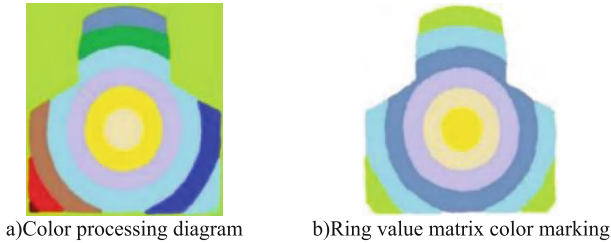


Fig. 8. Color tag

zones. Normalize this part and mark the areas with the same ring values on the target surface using the same color, as shown in Fig. 8b).

To facilitate a more intuitive observation of the hit outcomes, a human-computer interface for image processing and visualization was developed utilizing the PyQt5 [11] framework, as shown in Fig. 9. This interface displays the image of the target following current hit, along with pertinent data including the current number of rings, the highest and lowest ring counts, the total number of shots taken, and both the total and average number of rings achieved.

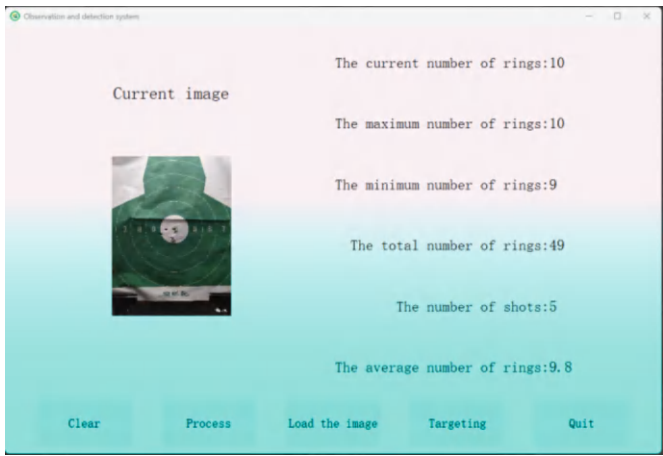


Fig. 9. Real-time target reporting interface

The primary functions of the system include the display of images, image processing, as well as the loading and clearing of images, in addition to presenting certain statistical information. By selecting the “Load” button, users can choose and upload an image, which will subsequently be displayed on the left side for the purpose of detecting target hits following the identification of holes.

In consideration of the influence of the above hole detection on the precision of the target, the scores of 50 images of the target surface were counted in this study, and all the hole ring values of each target surface were aggregated as the whole score of the target surface. The overall success rate for hole detection was 98.9%.

5 Conclusion

Based on the Siamese neural networks, the channel fusion Siamese neural network is designed to solve the difficult problem of hole location recognition and ring number reading, and a hole detection algorithm based on the network is proposed, which solves the problem that the hole is small and difficult to identify in the detection process of the target surface. At the same time, the network enhances its capacity to learn varying features through the adaptive allocation of channel weights. By integrating multiple channel fusion Siamese neural networks into the hole detection framework and determining the center coordinates of the holes, the research applied a ring zone calibration method to compute the ring value matrix, thereby achieving high-precision recognition of holes and accurate determination of ring values.

In order to more intuitively observe the shot results of shooting, the design of human-computer interaction interface shows how to use PyQt5 to create a GUI application for image display and processing. By adding image processing algorithm and more man-machine interaction function, the interface becomes a practical observation and detection tool.

The research on the target surface hole detection technology has a good application prospect in the practical life of automatic target reporting equipment in the future.

Acknowledgment. Shenyang University College Student Innovation Training Program Project Support + 202411035004.

References

1. Ijamaru, G.K., Nwajana, A.O., Oleka, E.U., et al.: Image processing system using MATLAB-based analytics. *Bull. Electr. Eng. Inform.* **10**(5), 2566–2577 (2021)
2. Serrano, N., Bellogín, A.: Siamese neural networks in recommendation. *Neural Comput. Appl.* **35**(19), 13941–13953 (2023)
3. Bölücü, N., Can, B., Artuner, H.: A siamese neural network for learning semantically-informed sentence embeddings. *Expert Syst. Appl.* **214**, 119103 (2023)
4. Li, X., et al.: Transformer-based visual segmentation: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 1–24 (2024)
5. Hussain, H., Tamizharasan, P.S., Rahul, C.S.: Design possibilities and challenges of DNN models: a review on the perspective of end devices. *Artif. Intell. Rev.* (2022). <https://doi.org/10.1007/s10462-022-10138-z>
6. Praveen, S.P., et al.: ResNet-32 and FastAI for diagnoses of ductal carcinoma from 2D tissue slides. *Sci. Rep.* **12**(1), 20804 (2022)
7. Thaufiq, P., Panyayot, C.: High performance and energy efficient Sobel edge detection. *Microprocess. Microsyst.* **87**, 104368 (2021)
8. Chen, W., Zou, B., Yang, J., et al.: The machined surface defect detection of improved super-pixel segmentation and two-level region aggregation based on machine vision. *J. Manuf. Process.* **80**, 287–301 (2022)
9. Yuewu, J., Yujin, Z., Jianxin, S.: Expression recognition based on key points and weight distribution residual network. *J. Comput. Eng. Appl.* **58**(17), 181–188 (2022)
10. Meihua, G., et al.: Multi-scale fusion grayscale algorithm for color images. *J. Comput. Eng. Appl.* **57**(4), 209–215 (2021)
11. Willman, J., Willman, J.: Overview of pyqt5. *Modern PyQt: Create GUI Applications for Project Management, Computer Vision, and Data Analysis*, pp. 1–42 (2021)



Intelligent Database Triggers Enable Advanced Analysis of Data Recorded in Audit Logs

Yongna Li^(✉), Cuiping Li, and Zhaoxia Cui

Binzhou Polytechnic, Binzhou 256600, Shandong, China
80320690@qq.com

Abstract. Traditional static database design accounting information system lacks a systematic audit trail mechanism, which makes it impossible to fully record data changes and operation history, and it is difficult to trace and analyze when problems occur. When using traditional data warehouses such as Oracle Data Warehouse, the data from multiple systems are incompatible when integrated, resulting in information islands in the audit process. This paper applies database triggers to automatically record and analyze operation logs to establish a more comprehensive and systematic audit trail mechanism. A more consistent and integrated audit mechanism is built so that information between different modules and systems can interact and form a complete audit chain. By designing an AuditLog audit log table, AFTER INSERT, AFTER UPDATE, and AFTER DELETE triggers in the database are created to write operation information into the audit log table. SQL query scripts are written to regularly analyze the data in the AuditLog table and count and filter abnormal operations. API interfaces are designed and ETL tools are used to integrate audit log data from different modules into a central database to achieve information sharing between modules. The research results show that after the implementation of the new audit trail mechanism, the integrity of the audit log is significantly improved, and the abnormal operation detection rate is as high as 1.07%, indicating that the system can identify and mark potential risk behaviors in real-time, enhancing the security and reliability of data. The user feedback survey results show that most users are satisfied with the system's log integrity and anomaly detection functions, especially in providing complete operation history and accurate anomaly marking. However, the system still needs to be further optimized in terms of response speed and performance to improve user experience.

Keywords: Accounting Information System · Audit Trail Mechanism · Database Triggers · Audit Log · Anomaly Detection

1 Introduction

With the rapid development of information technology, accounting information systems play an increasingly important role in enterprise management. However, traditional accounting information systems generally have the problem of imperfect audit trail mechanisms. When problems arise in the enterprise's finances, there is a lack of

effective tracing and analysis methods, which seriously affects the enterprise's decision-making efficiency and management risks. In addition, due to the incompatibility of audit trails in different systems, audit information islands are created, further increasing the complexity of the audit process.

This paper proposes to explore and improve the audit trail mechanism in the accounting information system to address the problems of audit deficiencies and information islands in the traditional static database design. To achieve this goal, the study adopts a combination of database triggers and ETL tools, designs a dedicated Audit Log table, and creates triggers such as AFTER INSERT, AFTER UPDATE, and AFTER DELETE to achieve automated audit log recording. SQL query scripts are also written for regular analysis of audit log data, and RESTful API interfaces are designed to facilitate the sharing of audit information between modules. Through these steps, the study successfully establishes a comprehensive, systematic, and efficient audit trail mechanism. The results of the study show that the integrity of the audit log and the detection rate of abnormal operations are significantly improved, and the feedback from users when using the system also shows a high level of satisfaction.

2 Related Works

In the research area of audit trails of accounting information systems, a number of researchers have provided relevant insights. Duggineni, Sasidhar proposed to manage and protect data by enhancing data integrity and security [1]. This study demonstrated that reliable data management can effectively address the risks of data elements and improve the efficiency of audit trails. Some scholars have explored the application of data warehouse technology in cross-system information integration [2], arguing that by constructing a centralized data platform, information silos can be effectively broken and the overall efficiency of auditing can be improved. These studies provide a theoretical basis for the improvement of audit trail mechanisms, highlighting the key role of compatibility in the audit process. The current research still has some shortcomings, especially in how to realize effective integration and information sharing among different systems still lacks systematic and operable solutions. Therefore, it has become an important research direction to further study how to improve the overall skills of the audit trail mechanism through more relevant modern technological means. This paper builds on this foundation by exploring how to combine database triggers with data integration tools to create a more comprehensive and compatible audit trail mechanism.

To address the shortcomings of current research, some researchers have focused on the application of database triggers and structured query language in the audit trail mechanism. Togatorop Parmonangan pointed out that the use of database triggers can automatically record every data change, effectively improving the integrity and timeliness of audit trail [3]. The advantage of this method is that it can be automatically triggered when data is added, deleted, modified, or checked, ensuring that the audit log can accurately reflect each operation and provide detailed audit evidence. Although the use of triggers improves the efficiency of audit trail, it still faces the problem of isolated audit log information. This isolation limits the sharing and compatibility of information between different modules and systems, and hinders the formation of a comprehensive

audit chain. To solve this problem, this paper combines database triggers with ETL tools [4] to propose a more comprehensive audit trail mechanism. Through data integration and coordination between modules, effective information flow is achieved; the island phenomenon between different systems is eliminated; the efficiency and reliability of the entire audit process are improved.

3 Methods

3.1 Design of AuditLog Audit Log Table

In building an audit trail mechanism, it is necessary to design a dedicated AuditLog audit log table to systematically store all relevant audit information. The design of the audit log table is to ensure comprehensive recording of data changes and subsequent efficient analysis. The audit log table is created using SQL DDL statements [5]. The table structure contains the fields LogID, OperationType, Timestamp, UserID, and RecordID. Among them, LogID is the primary key, which is an automatically incremented integer that uniquely identifies the audit record [6]. OperationType records the three operation types of INSRET, UPDATE, and DELETE to facilitate data classification and analysis. Timestamp uses the database's built-in CURRENT_TIMESTAMP [7] to record the operation time. UserID is the user identity who performs the operation to ensure the traceability of the operation responsibility. RecordID is the unique identifier of the operation record, which facilitates the association of audit information with specific business data. The data table is shown in Table 1.

Table 1. Audit log table structure

Column Name	Data Type	Constraints	Description
LogID	Int	PRIMARY KEY, AUTO_INCREMENT	Unique identifier for each audit record
Operation Type	Varchar(10)	NOT NULL	Type of operation (INSERT, UPDATE, DELETE)
Timestamp	DATETIME	DEFAULT CURRENT_TIMESTAMP	Exact time of the operation
UserID	Varchar(50)	NOT NULL	Identity of the user performing the operation
RecordID	Varchar(50)	NOT NULL	Unique identifier of the record being operated on

4 To ensure data integrity and consistency, non-null constraints are added to the UserID and RecordID fields to prevent the insertion of invalid data. Check constraints are set on the OperationType field to ensure that only valid operation types are allowed. To improve query efficiency, indexes are created for the Timestamp and UseID fields [8] to facilitate

quick retrieval of operation records within a specific time range or for a specific user, effectively improving the system’s processing speed for large amounts of data during the audit process. Since the audit log table may grow rapidly, a data cleanup mechanism [9] is added. The time threshold is set to save audit records for the past three years, and expired records are archived and deleted regularly to ensure database performance and effective use of storage space.

3.2 Creation of Database Triggers

To achieve automated audit records, database triggers need to be created in the target table. When data operations occur, information is automatically written to the audit log table. Triggers are created for different operation types in the database. The AFTER INSERT trigger is triggered when new data is inserted. The NEW keyword is used to obtain the newly inserted record information. The operation information is written to the audit log table for automatic recording. The AFTER UPDATE trigger is triggered when data is updated. The OLD keyword is used to obtain the original state of the updated record and record the change information, which helps track historical changes. The AFTER DELETE trigger is triggered when data is deleted. The ID of the deleted record is recorded so that auditors can track which data is deleted, providing important basis for subsequent analysis. Figure 1 shows the framework of the detection model.

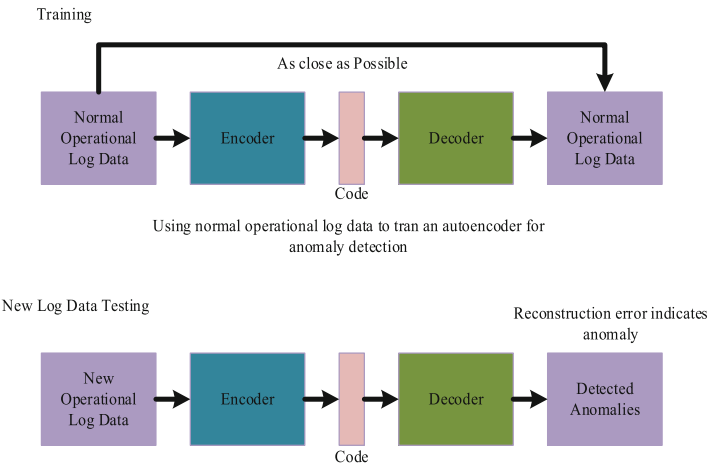


Fig. 1. Anomaly detection model framework

The system extracts operation records from the audit log, including operation type, user ID, timestamp, etc. The encoder and decoder compress the high-dimensional data and restore it to data similar to the input data. Through the set rules and thresholds, the restored data is analyzed to mark abnormal operations. By implementing the abnormal operation detection mechanism, the system can identify and mark potential risk behaviors in real-time, improving data security.

After creating a trigger, it is necessary to conduct sufficient testing to ensure that it functions properly. Insert, modify, and delete operations are performed on the target

table, and the AuditLog table is queried to check whether the corresponding operation information is successfully recorded. Whether the record includes the operation type, user ID, and record ID is checked to ensure that the trigger works properly. An exception handling mechanism is added to the trigger to ensure that when invalid data is inserted, the trigger can capture and record errors such as data integrity errors and data type mismatches. This processing mechanism improves the robustness of the trigger and ensures the integrity of data operations. The creation of triggers can record all operations on data tables in real-time, solving problems missed by traditional methods, improving the degree of audit automation and data traceability, and enhancing the overall reliability of the accounting information system.

3.3 Writing SQL Query Scripts

In the audit trail mechanism, SQL query scripts [10] are used to analyze and filter audit log data. The required information is obtained from the log table by writing data extraction and summarization scripts and the type and frequency of operations are summarized with the help of GROUP BY and COUNT functions. The filter fields include operation time, type, user ID, and details. The CASE statement is used to judge the abnormal situation of specific operations to identify potential abnormal operations.

To ensure timely analysis of audit data, SQL query scripts are written to be incorporated into the timed tasks of the database, and SQL Agent is used to set up automatic campaigns in the early hours of each day to ensure continuity and timeliness of data analysis. After the script is executed, the data results are exported to a CSV file for further analysis and writing by the audit team. The SQL query script can not only effectively extract and analyze the data in the AuditLog table, but also identify abnormal operations in a timely manner, helping the audit team to quickly locate and analyze the root cause of the event when a problem occurs, ensuring the effectiveness of the audit trail mechanism and the integrity of the data, and improving the security and reliability of the system.

3.4 Integration of API Interface and ETL Tool

Integrating API interface and ETL tool is the key to realize the sharing of different modules and log data. Designing RESTful API interface makes it easier for different systems to interact with AuditLog table through HTTP request. The main function of API is to create GET request interface [11], allowing external system to audit log data as required. Then, designing POST request interface enables other modules to directly write audit information into AuditLog table when generating relevant operations. Through these two API interfaces, real-time interaction of audit information between modules is realized, improving the accuracy and timeliness of data.

Appropriate ETL tool is selected. This paper adopts Apache Nifi [12] to extract, transform and load data from different systems into central database. ETL tool is configured to extract audit log data from each business system regularly, and connectors and custom query scripts are used to obtain required information. After data extraction, necessary data cleaning and transformation are performed, and time format, field name, etc., are unified to ensure data format consistency, facilitating subsequent analysis and use. The cleaned data is loaded into AuditLog table in central database regularly, and

daily scheduled tasks are set to automatically update data and ensure the latest status of data. During the integration process, strict access control policies are set to ensure data security and privacy. Permission management is implemented through the API interface to ensure that only authorized users can access audit log data. At the same time, the permission settings of the ETL tool are used to limit the scope and method of data access.

After the integration is completed, the execution effect of the API interface and the ETL process is verified regularly. Monitoring tools such as Prometheus [13] are used to track the number of API calls, response time [14] and the success rate of ETL tasks to ensure stable operation of the system. Through log analysis, potential problems can be discovered and resolved in a timely manner. The integration of the API interface and the ETL tool [15] realizes the efficient sharing of audit log data between different modules, solves the problem of information islands, and improves the overall effectiveness and reliability of the audit trail mechanism.

4 Results and Discussion

4.1 Log Integrity

In accounting information systems, the integrity of audit logs is an important basis for ensuring data traceability, compliance and security. In order to quantify the integrity of audit logs, the effective record ratio is used to reflect the situation of effective records in total records. The effective record ratio formula is:

$$C = \frac{R_{\text{valid}}}{R_{\text{total}}} \times 100\% \quad (1)$$

R_{valid} is the number of effective records, and R_{total} is the total number of records. Audit logs are regularly reviewed and monitored at various stages of system operation. During the audit process, special attention is paid to the handling of abnormal records. When the number of abnormal records exceeds the preset threshold, the system automatically triggers an alarm to remind the relevant responsible persons to handle them in time. The types of abnormal records are analyzed, and data input and processing processes are improved in a targeted manner to reduce the probability of errors. To prevent data loss or damage, a data backup mechanism is established to export and encrypt audit logs on a monthly basis. The backup data is strictly controlled to ensure that sensitive information is not leaked. The integrity of the audit log is quantified. The change trend of the integrity score in each quarter is illustrated in Fig. 2.

The integrity score fluctuates relatively little, indicating that the integrity of the audit log is effectively maintained. Regular auditing and monitoring ensure the validity of the logs, and the rapid processing of abnormal records and strict security control measures enhance the security of the system. Data backup and recovery strategies provide guarantees for the long-term storage of logs. In summary, the integrity of the audit log provides solid support for the compliance and security of the accounting information system, laying the foundation for subsequent audits and analysis.

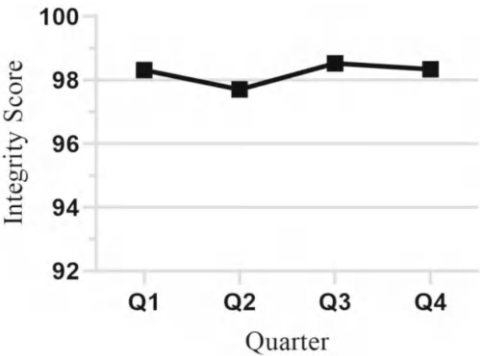


Fig. 2. Change trend of log integrity scores

4.2 Abnormal Operation Detection Rate

Abnormal operation detection improves the data security and integrity of the system, identifies abnormal behavior caused by human error, malicious attack or system failure, and prevents potential financial risks and data loss. The abnormal behaviors studied in this paper include repeated operations in a short period of time, financial changes of abnormal amounts, requests for operations beyond authority, and insertion of records that do not conform to the data format. Detection methods include statistical analysis, threshold monitoring, and user behavior analysis. Statistical analysis uses SQL scripts to regularly count the frequency of operations to identify anomalies. Threshold monitoring sets a reasonable range. When the operation exceeds the threshold, it is marked as abnormal. User behavior analysis judges deviations based on historical operations. The abnormal operation detection rate calculation formula is:

$$DR = \frac{N_{\text{anomalies}}}{N_{\text{total}}} \times 100\% \tag{2}$$

$N_{\text{anomalies}}$ is the number of abnormal operations detected, and N_{total} is the total number of operations. Formula 2 is used to quantify the detection effect of abnormal operations to evaluate the effectiveness of the audit trail mechanism. Table 2 shows the abnormal operation detection rate in different quarters, reflecting the system’s prosecution effect on abnormal operations in various time periods.

Table 2. Abnormal operation detection rate in each time period

Quarter	Total Operations	Abnormal Detected Anomalies	Abnormal Detection Rate (%)
Q1	12000	118	0.98
Q2	15000	92	0.61
Q3	14300	153	1.07
Q4	13400	84	0.63

As can be seen from Table 2, the detection rate in Q3 is the highest, reaching 1.07%. The analysis of the reasons shows that a new round of system monitoring and user training is implemented during Q3, which improves users’ awareness of abnormal operations and self-prevention awareness. The detection rates in Q2 and Q4 are lower, which is related to the reduction in operation frequency and the adjustment of system monitoring parameters.

4.3 User Feedback and Satisfaction

This paper designs an automatic audit log recording and analysis mechanism based on SQL and database triggers. In order to evaluate the effectiveness of the system in practical applications, the practicality and user experience of the system are analyzed through user feedback and satisfaction surveys. At the beginning of the system operation, feedback from auditors and other users of the accounting information system who use the audit trail mechanism is collected. The survey includes whether auditors can obtain complete operation history and data change information from the audit log, the accuracy and timeliness of the system in marking and reporting abnormal operations, the user’s experience in extracting audit logs and querying specific operation records, and the ability of cross-module data integration functions to reduce information islands. Through questionnaires and interviews, the user feedback collected shows that most users have a

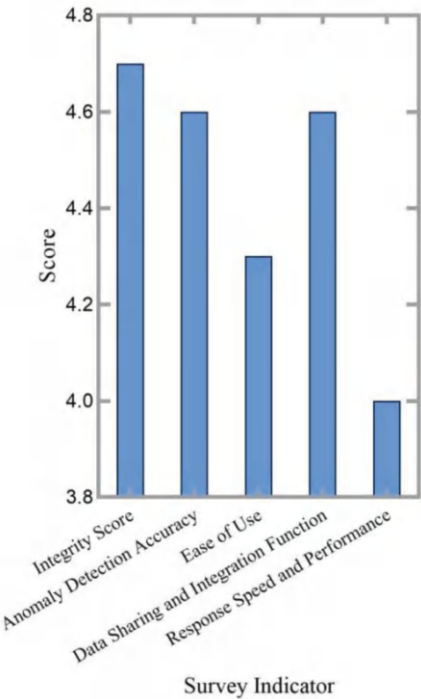


Fig. 3. Survey satisfaction scores

positive attitude towards the overall function of the system, especially the performance of the system in providing complete logging and anomaly detection. Figure 3 shows the user satisfaction scores for different survey indicators.

As can be seen from Fig. 3, the log integrity score is the highest, indicating that users are very satisfied with the system's log integrity function. This means that the system performs well in data integrity and reliability and can fully meet users' needs for records. The response speed and performance scores are relatively low, indicating that the system's response speed and performance are not good enough, which may affect the user's overall experience.

5 Conclusions

This paper discusses the use of SQL and database triggers to realize the automated audit trail mechanism of accounting information systems. By designing the AuditLog table and triggers, the system can record data changes completely and in real-time. SQL queries and ETL tools are combined to achieve data integration and sharing, solving the problem of information islands. The research results show that this method significantly improves the integrity and anomaly detection rate of audit logs, and user feedback is positive. However, the system needs to be optimized in terms of response speed and performance. Future research can focus on optimizing SQL scripts and API interfaces to improve performance and user experience, and ensure the effectiveness and reliability of audit trail.

References

1. Duggineni, S.: Impact of controls on data integrity and information systems. *Sci. Technol.* **13**(2), 29–35 (2023)
2. Al-Okaily, A., et al.: An empirical study on data warehouse systems effectiveness: the case of Jordanian banks in the business intelligence era. *EuroMed J. Bus.* **18**(4), 489–510 (2022)
3. Togatorop, P., et al.: Database audit system design and implementation. *Jurnal Mantik* **5**(4), 2535–2541 (2022)
4. Qaiser, A., et al.: Comparative analysis of ETL tools in big data analytics. *Pak. J. Eng. Technol.* **6**(1), 7–12 (2023)
5. Nurmatovich, T.I., Azizjonog, N.A.: The SQL server language and its structure. *Am. J. Open Univ. Educ.* **1**(1), 11–15 (2024)
6. Kossmann, J., Papenbrock, T., Naumann, F.: Data dependencies for query optimization: a survey. *VLDB J.* **31**(1), 1–22 (2022)
7. Ma, R., et al.: Modeling and querying temporal RDF knowledge graphs with relational databases. *J. Intell. Inf. Syst.* **61**(2), 569–609 (2023)
8. Amini, M., Rahmani, A.: Agricultural databases evaluation with machine learning procedure. *Aust. J. Eng. Appl. Sci.* **8**(2023), 39–50 (2023)
9. Borrohou, S., Fissoune, R., Badir, H.: Data cleaning survey and challenges—improving outlier detection algorithm in machine learning. *J. Smart Cities Soc.* **2**(3), 125–140 (2023)
10. Tojiboyev, N., et al.: Basics of SQL for audit data retrieval and analysis. *J. Emerg. Technol. Account.* **19**(1), 237–265 (2022)

11. Olabanji, S.O.: Advancing cloud technology security: Leveraging high-level coding languages like Python and SQL for strengthening security systems and automating top control processes. *J. Sci. Res. Rep.* **29**(9), 42–54 (2023)
12. Carthen, C.D., et al.: A novel spatial data pipeline for orchestrating apache NiFi/MiNiFi. *Int. J. Softw. Innov. (IJSI)* **12**(1), 1–14 (2024)
13. Jani, Y.: Unified monitoring for microservices: implementing prometheus and grafana for scalable solutions. *J. Artif. Intell. Mach. Learn. Data Sci.* **2**(1), 848–852 (2024)
14. Khriji, S., et al.: Design and implementation of a cloud-based event-driven architecture for real-time data processing in wireless sensor networks. *J. Supercomput.* **78**(3), 3374–3401 (2022)
15. Gupta, A., et al.: The role of managed ETL platforms in reducing data integration time and improving user satisfaction. *J. Res. Appl. Sci. Biotechnol.* **1**(1), 83–92 (2022)



Improvement of Principal Component Analysis Algorithm and Its Simulation Experiment

Ling Zhang^(✉)

Liaoning Geology Engineering Vocational College, Dandong, China
442623405@qq.com

Abstract. As an important branch of mathematical statistics, multivariate statistical analysis has developed rapidly, its theory is more rigorous, its content is more solid, and it has a wide range of practical application value. As one of the dimensionality reduction techniques of multivariate statistical analysis, principal component analysis transforms many highly correlated variables into mutually independent or unrelated variables, simplifies the multi-variable high-dimensional space problem into a low-dimensional comprehensive index problem, and reflects the information of the original variables as much as possible through the comprehensive variables. In this paper, a mathematical model of principal component analysis (PCA) is constructed, and some improvement measures are proposed in terms of processing raw data, KMO test and Bartlett spherical test. Laboratory safety evaluation indicators were used to collect original data by expert scoring method, and Matlab software was used to simulate the improved algorithm proposed in this paper. The original 20 evaluation indicators were replaced by 4 main factors, and factors replaced the original variables to participate in data modeling, so as to overcome the defects caused by too many variables in the analysis process.

Keywords: Principal Component Analysis · Algorithm Improvement · Simulation Experiment · Mathematical Model · KMO Test · Bartlett Spherical Test

1 Introduction

Scientific research is a process of repeated analysis, in which explanations related to certain social or natural phenomena are usually selected as targets, and these objectives are tested through data collection and analysis, and improved explanations are proposed for the phenomena after the data collected through experiments or observations are analyzed. In this process of repeated learning, some variables are often added to the study and some are removed. Therefore, due to the complexity of most phenomena, researchers are required to collect observations of many different variables, and this type of method is called multivariate statistical analysis. The research of multivariate statistical analysis method has practical application value [1, 2]. Principal component analysis transforms many highly correlated variables into mutually independent or unrelated variables, simplifies the multi-variable high-dimensional space problem into a low-dimensional comprehensive index problem, and reflects the information of the original variables as much

as possible through the comprehensive variables. As a method with the same dimensionality reduction effect as factor analysis in multivariate statistical analysis, principal component analysis can not only analyze the correlation between indicators, but also reduce the dimensionality of large sample data. It can not only simplify the number of indicators, but also avoid the complexity and redundancy of indicators that have little impact on the final evaluation result when analyzing large sample data. It is also conducive to the analysis of some large sample data, eliminating the indicators that have little impact on the evaluation results and retaining the indicators that have a greater impact on the evaluation results, so as to have a deeper understanding of the important factors that affect the evaluation indicators and better quantitative analysis according to the information extraction value of the indicators. The method based on principal component analysis makes the evaluation results of large sample data more accurate and more objective. Because of the flexibility of the results of principal component comprehensive evaluation, one of the dimensionality reduction methods in multivariate statistical analysis has more important research value.

2 Related Works

As one of the dimensionality reduction techniques of multivariate statistical analysis, principal component analysis has the outstanding effect of increasing the ratio of sample size to the size of observations [3]. From the perspective of literature review, domestic and foreign scholars have carried out multi-angle research [4, 5]: improving the principal component analysis method from the aspects of principal component number selection and principal component rotation, introducing the factor rotation idea to principal component analysis, and achieving a good improvement analysis effect. The inertia weight of entropy weight method is used as the comprehensive weight of principal component analysis, instead of automatically generating the weight in the original principal component analysis process, and a good dimensionality reduction effect is obtained, which overcomes the problem that the cumulative contribution rate of the first principal component is not enough, and it is difficult to determine the index weight when selecting multiple indicators for analysis. In order to make the dimensionality reduction effect of the index data obvious, the dependent variable of the sample data is used as the premise, the prediction equation of principal component analysis is established by comparing the results, and the model is optimized to achieve a good dimensionality reduction effect. On the basis of finding out the robust model of principal component analysis, statistical researchers give an independent definition of principal component analysis, and introduce nonlinear thinking into principal component analysis, and put forward a nonlinear principal component analysis method. Starting from the sample analysis matrix, the conventional principal component analysis usually uses the correlation coefficient matrix or covariance matrix of the principal component sample for comprehensive evaluation, and the conventional covariance matrix is usually very sensitive to the sample's inferior point value, resulting in poor stability of the analysis results. By improving the covariance matrix, the stability of principal component analysis method is improved.

Predecessors have conducted in-depth research on principal component analysis algorithm, but there are still some problems to be solved [6–8]: The selection of eigenvector direction is still an unsolved problem, and some scholars have put forward their

own views on determining the direction of eigenvector. In the analysis of specific problems, there are different selection principles, but how to choose in practical application is still a research hotspot. For the interpretation of principal components, many scholars make explanations according to the magnitude of positive vectors, and some scholars study explanations according to positive and negative direction vectors, but the interpretation of principal components still needs in-depth research. Some scholars regard the weight selection of comprehensive evaluation as the standard, and some scholars regard entropy as the weight selection, but the rationality of weight still needs further discussion. Factor rotation problem, on the basis of principal component analysis, many scholars study the same factor rotation as factor analysis. Whether the principal component is suitable for rotation and whether the solution of the principal component after rotation is reasonable still need further research. According to the previous research results, the appropriate number of eigenvalue selection is greater than or equal to the appropriate problem. However, when applied to specific problems, many important measurement indicators are not included in this scope, and further selection needs to be based on specific problems. At present, there are many researches on the pretreatment of raw indicators. How to deal with raw indicators according to specific problems and how to choose processing methods are still the key contents of comprehensive evaluation by principal component analysis method.

In this paper, the principal component analysis algorithm is improved, and the laboratory safety evaluation data is used for simulation experiments. Laboratory safety involves many aspects such as personal life, water and electricity, equipment and equipment, hazardous waste disposal and property anti-theft. The frequent occurrence of laboratory safety accidents is due to the indifferent safety awareness, the simple hardware environment, the lagging security system, and the scientific research pressure that is always rushing to produce results. To prevent the occurrence of laboratory safety accidents and ensure the smooth development of scientific experiments is an urgent problem to be solved. Laboratory safety assessment is an important part of comprehensive laboratory management, and many studies have been carried out by predecessors [9, 10]. AHP and CRITIC algorithm are used to calculate subjective and objective weights, and a comprehensive laboratory safety evaluation model based on combined weighted rank sum ratio is constructed, which can provide data support for laboratory safety management, and be widely applied to laboratory safety work evaluation and laboratory safety management assessment. DHGF algorithm has strong applicability and operability, and can ensure the scientific and accuracy of the entire evaluation results. The evaluation results have been unanimously recognized by the experts of the laboratory safety Committee. In the future, this method can be used to further systematically evaluate the safety of the laboratory environment. According to the analysis of the evaluation results, the direction of further improvement of laboratory safety management is given, which has reference significance for the prevention of laboratory accidents in universities.

3 Methods

In this section, the principal component analysis algorithm is studied first, and then the principal component analysis algorithm is improved.

3.1 Principal Component Analysis Algorithm

3.1.1 Definition of Principal Components

Let the sample standard deviation of random variable $X = [X_1, X_2, \dots, X_p]$ be:

$$D(X) = [D(X_1), D(X_2), \dots, D(X_p)] \quad (1)$$

First make the standard transformation:

$$C_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ip}x_p \quad (i = 1, 2, \dots, p) \quad (2)$$

If $C_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$, and $\text{Var}(C_1)$ is the largest, C_1 is said to be the first principal component;

If $C_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$, $(a_{21}, a_{22}, \dots, a_{2p})$ is perpendicular to $(a_{11}, a_{12}, \dots, a_{1p})$ and $\text{Var}(C_2)$ is the largest, C_2 is said to be the second principal component;

Similarly, there can be a third, fourth, fifth... There are at most p principal components.

3.1.2 Properties of Principal Components

Property 1. n distinct feature roots are denoted $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$, and the sum of variance of X_1, X_2, \dots, X_n is equal to the sum of feature roots:

$$\sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^n \lambda_i, \quad \sigma_i^2 = E(X_i^2) \quad (i = 1, 2, \dots, n) \quad (3)$$

Property 2. Let $\rho_{ij} = \rho(X_i, X_j)$, $\sigma_i^2 = E(X_i^2)$, β_{ij} is the j -th component of the eigenvector β_i of λ_i , then:

$$\rho_{ij} = \frac{\sqrt{\lambda_i} \beta_{ij}}{\sigma_j} \quad (i, j = 1, 2, \dots, n) \quad (4)$$

Property 3. The variance of each principal component is decreasing successively, that is:

$$\text{Var}(C_1) \geq \text{Var}(C_2) \geq \dots \geq \text{Var}(C_p) \quad (5)$$

Property 4. The total variance does not increase or decrease, that is:

$$\begin{aligned} &\text{Var}(C_1) + \text{Var}(C_2) + \dots + \text{Var}(C_p) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_p) = p \end{aligned} \quad (6)$$

3.1.3 Mathematical Model of Principal Component Analysis

All sample values are represented by matrix as follows:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \quad (7)$$

Standardize the above formula with the following formula:

$$x_{ij} = \left(y_{ij} - \frac{1}{m} \sum_{i=1}^m y_{ij} \right) / \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{ij} - \mu_j)^2} \quad (8)$$

After standardization, the indexes of different dimensions have comparability. The data matrix after standardized processing is as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (9)$$

The correlation coefficient is as follows:

$$r_{ij} = \frac{1}{m-1} \sum_{k=1}^m X_{ik} X_{jk} \quad (i, j = 1, 2, \cdots, n) \quad (10)$$

The correlation matrix composed of correlation coefficients is expressed as:

$$R = \frac{1}{m-1} X'X = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \quad (11)$$

The characteristic polynomial is represented as follows:

$$\begin{aligned} |\lambda E - R| &= \begin{vmatrix} \lambda - r_{11} & -r_{12} & \cdots & -r_{1n} \\ -r_{21} & \lambda - r_{22} & \cdots & -r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ -r_{n1} & -r_{n2} & \cdots & \lambda - r_{nn} \end{vmatrix} \\ &= \lambda^n + r_1 \lambda^{n-1} + r_2 \lambda^{n-2} + \cdots + r_{n-1} \lambda + r_n \end{aligned} \quad (12)$$

The eigenvector matrix U with p -dimension can be decomposed into:

$$U = [U_1 U_2 \cdots U_k U_{k+1} U_{k+2}] = \begin{bmatrix} U(1) & U(2) \\ n \times k & n \times (n-k) \end{bmatrix} \quad (13)$$

Substituting U into the basic equations of factor analysis are:

$$\begin{aligned}
 x_{n \times m} &= U_{n \times n} f_{n \times m} = [U_1 U_2] \begin{bmatrix} f(1) \\ f(2) \end{bmatrix} \\
 &= U(1)_{n \times k} f(1)_{k \times m} + U(2)_{n \times (n-k)} f(2)_{(n-k) \times m} \\
 &= U(1) f(1) + e
 \end{aligned} \tag{14}$$

Principal component analysis expression:

$$\begin{cases} F_1 = u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n \\ F_2 = u_{21}x_1 + u_{22}x_2 + \cdots + u_{2n}x_n \\ \cdots \\ F_p = u_{p1}x_1 + u_{p2}x_2 + \cdots + u_{pn}x_n \end{cases} \tag{15}$$

3.2 Principal Component Analysis Algorithm Improvement

The principal component analysis algorithm is improved from many angles to improve the effectiveness of the analysis.

3.2.1 Processing Raw Data

According to the original model of principal component analysis, in the sample of n objects to be evaluated, m evaluation indicators are usually used to describe each evaluation object, and the following definition is given.

Definition 1. $X = \{x_1, x_2, \cdots, x_m\}$, m evaluation objects are called the original indicators, $\bar{x} = \frac{1}{m} \sum_{k=1}^m x_i$ is the mean value of the original indicators, and the covariance of the original indicators is:

$$V = \frac{1}{m-1} \sum_{k=1}^m (x_i - \bar{x}_i)(x_j - \bar{x}_j) \tag{16}$$

Definition 2. $\omega = \{\omega_1, \omega_2, \cdots, \omega_m\}$ is the inertia coefficient corresponding to the original indicator, where:

$$\omega_{ij} = \lambda_{ij} \frac{v_i}{\sum_{i=1}^m v_i} + (1 - \lambda_{ij}) \frac{g_i}{\sum_{i=1}^m g_i} \tag{17}$$

In the above equation, $i = 1, 2, \cdots, n, j = 1, 2, \cdots, m$.

Definition 3. $T = \{\omega_1 x_1, \omega_2 x_2, \dots, \omega_m x_m\}$ is the first-level optimization index, and the mean value of the first-level optimization index is:

$$\tilde{x} = \frac{1}{m} \sum_{k=1}^m \omega_k x_k \quad (18)$$

The first-level optimization index covariance is:

$$\bar{V} = \frac{1}{m-1} \sum_{k=1}^m (\omega_{ki} x_i - \tilde{x}_i)(\omega_{kj} x_j - \tilde{x}_j) \quad (19)$$

Definition 4. $Y = \{y_1, y_2, \dots, y_m\}$ is the second-level optimization index, and $y_{ij} = \frac{\omega_{ij} x_{ij}}{\tilde{x}}$, the covariance of the second-level optimization index is:

$$\begin{aligned} \tilde{V} &= \frac{1}{m-1} \sum_{k=1}^m (y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j) \\ &= \frac{1}{m-1} \sum_{k=1}^m \frac{(\omega_{ki} x_i - \tilde{x}_i)}{\tilde{x}_i} \cdot \frac{(\omega_{kj} x_j - \tilde{x}_j)}{\tilde{x}_j} \\ &= \frac{1}{\tilde{x}_i \tilde{x}_j} \cdot \frac{1}{m-1} \sum_{k=1}^m (\omega_{ki} x_i - \tilde{x}_i) \cdot (\omega_{kj} x_j - \tilde{x}_j) \\ &= \frac{1}{\tilde{x}_i \tilde{x}_j} \bar{V} \end{aligned} \quad (20)$$

In the above definition, by assigning an inertia coefficient to the original index for weighting processing, the original index is converted into the defined first-level optimization index, and then the converted first-level optimization index is converted into the second-level optimization index by averaging.

3.2.2 KMO Test

Let $(X_i, Y_i)(i = 1, 2, \dots, n)$ be the sample taken from the population, and the mean of $X_i(i = 1, 2, \dots, n)$ is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (21)$$

The mean of $Y_i(i = 1, 2, \dots, n)$ is:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (22)$$

The simple linear correlation coefficient of $(X_i, Y_i)(i = 1, 2, \dots, n)$ is:

$$\rho = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (23)$$

Let, k is the number of fixed variables; z_1, z_2, \dots, z_k is the fixed variable; r_{xy} is a simple correlation coefficient between the variables x and y .

When $k = 1$, the formula for calculating the partial correlation coefficient is as follows:

$$r_{xy, z_1} = \frac{r_{xy} - r_{xz_1} r_{yz_1}}{\sqrt{(1 - r_{xz_1}^2)(1 - r_{yz_1}^2)}} \quad (24)$$

When $k \geq 2$, the formula for calculating the partial correlation coefficient is as follows:

$$r_{xy, z_1 z_2 \dots z_k} = \frac{r_{xy, z_1 z_2 \dots z_{k-1}} - r_{xz_k, z_1 z_2 \dots z_{k-1}} r_{yz_k, z_1 z_2 \dots z_{k-1}}}{\sqrt{(1 - r_{xz_k, z_1 z_2 \dots z_{k-1}}^2)(1 - r_{yz_k, z_1 z_2 \dots z_{k-1}}^2)}} \quad (25)$$

The calculation formula of KMO test statistics is as follows:

$$KMO = \frac{P}{P + R} \quad (26)$$

The judging criteria of KMO index values are shown in Table 1.

Table 1. Standard KMO index

No	KMO value	Discriminate declaration	Applicability
1	0.6437	Excellent for analysis	Excellent
2	0.6124	Suitable for analysis	Good
3	0.5362	Can be analyzed	Middling
4	0.4778	Barely enough to analyze	Common
5	0.4445	Not suitable for analysis	No good
6	0.4235	Very unsuitable for analysis	Unacceptable

3.2.3 Bartlett Sphericity Test

The statistical calculation formula of Bartlett sphericity test is as follows:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \cdot \ln(|\det(R)|) \quad (27)$$

In the above formula, n is the sample size, p is the number of variables, R is the correlation coefficient matrix between variables, and $\det(R)$ is the determinant of the correlation coefficient matrix.

4 Results and Discussion

The simulation experiment is to verify the improved principal component analysis algorithm by using the laboratory safety evaluation index.

4.1 Principal Component Analysis Environment

The simulation experiment of this study uses Matlab software, efficient programming environment and rich application libraries, which makes the algorithm design and simulation experiment simple and efficient, and the third-party toolbox can realize various specific functions. In addition to Matlab software, other hardware and software support is also required. The experimental environment is shown in Table 2.

Table 2. Experimental environment for improving principal component analysis algorithm

No	Name	Model or version
1	Simulation software	Matlab 7.0
2	Auxiliary software	Excel 2017
3	Computer	Dell Vostro3020SFF
4	Operating system	Windows 11 Home Basic 64bit
5	CPU	Intel Core i5 13400 2.5 GHz
6	Memory	DDR4 3200 MHz 16 GB
7	Data interface	2×USB2.0,2×USB3.1 Type-A
8	Desktop management	360Assistant 11.0.0.2051

4.2 Principal Component Analysis Index

The principal component analysis index uses the laboratory safety evaluation index. Based on the research and analysis of relevant literature on laboratory safety evaluation at home and abroad, and combined with the practical needs of laboratory safety management in the new era, the laboratory safety evaluation index system was established, as shown in Table 3.

4.3 Results and Discussion of Principal Component Analysis

Expert scoring method was adopted for data acquisition. 7 experts were invited to score each indicator of 15 laboratories respectively. For each indicator, the highest and lowest scores are removed, and the remaining 5 scores are averaged as a scoring result.

Table 3. Laboratory safety evaluation index system

No	Code	Index name
1	x_1	Standard operating procedure
2	x_2	Safety management system
3	x_3	Safety management organization
4	x_4	Safety management responsibility
5	x_5	Emergency equipment management
6	x_6	Emergency identification management
7	x_7	First aid materials management
8	x_8	First aid measures plan
9	x_9	General equipment management
10	x_{10}	Special equipment management
11	x_{11}	Protective equipment management
12	x_{12}	Instrument safety management
13	x_{13}	Overall spatial layout
14	x_{14}	Security identifier setting
15	x_{15}	Clear evacuation route
16	x_{16}	Disposal of waste
17	x_{17}	Access control management system
18	x_{18}	Video monitoring system
19	x_{19}	Visual identification of hazards
20	x_{20}	Identification of unsafe behavior

4.3.1 Eigenvalue and Variance Contribution Rate

The calculation results of eigenvalues and variance contribution rates are shown in Table 4.

As can be seen from the above table, the cumulative variance contribution rate of the first four factors accounts for 98.372% of the total variance, which is greater than 90%, indicating that these four factors can summarize most of the information of the original data.

4.3.2 Factor Load Matrix

The calculation results are shown in Table 5.

Table 4. Eigenvalue and variance contribution rate

Component	Total	Variance %	Accumulate %
1	1635.157	71.253	71.253
2	339.780	14.806	86.059
3	151.962	6.622	92.681
4	130.594	5.691	98.372
5	30.297	1.320	99.692
6	3.926	0.171	99.863
7	0.995	0.043	99.906
8	0.771	0.034	99.940
9	0.575	0.025	99.965
10	0.422	0.018	99.983
11	0.199	0.009	99.992
12	0.104	0.005	99.997
13	0.052	0.002	99.999
14	0.026	0.001	100.000
15	3.770E−14	1.643E−15	100.000
16	1.700E−14	7.408E−16	100.000
17	1.283E−14	5.590E−16	100.000
18	−9.955E−15	−4.338E−16	100.000
19	−1.920E−14	−8.368E−16	100.000
20	−4.176E−14	−1.820E−16	100.000

4.3.3 Principal Component Analysis Expression

Principal components are represented by f_1, f_2, f_3, f_4 respectively, then the expression of principal component analysis is as follows:

$$\begin{aligned} f_1 = & 0.069x_1 + 0.069x_2 + 0.060x_3 + 0.062x_4 \\ & + 0.058x_5 + 0.058x_6 + 0.059x_7 + 0.056x_8 \\ & + 0.040x_9 + 0.037x_{10} + 0.040x_{11} + 0.040x_{12} \\ & + 0.074x_{13} + 0.072x_{14} + 0.061x_{15} + 0.074x_{16} \\ & + 0.066x_{17} + 0.062x_{18} + 0.063x_{19} + 0.062x_{20} \\ f_2 = & 0.180x_1 + 0.157x_2 + 0.166x_3 + 0.182x_4 \\ & + 0.063x_5 + 0.073x_6 + 0.068x_7 + 0.060x_8 \\ & + 0.076x_9 + 0.073x_{10} + 0.073x_{11} + 0.074x_{12} \\ & - 0.152x_{13} - 0.144x_{14} - 0.132x_{15} - 0.146x_{16} \end{aligned}$$

Table 5. Factor load matrix

Index	component			
	1	2	3	4
x_1	0.069	0.180	-0.059	0.017
x_2	0.069	0.157	-0.085	-0.006
x_3	0.060	0.166	-0.084	-0.028
x_4	0.062	0.182	-0.095	-0.039
x_5	0.058	0.063	-0.167	-0.111
x_6	0.058	0.073	-0.186	-0.135
x_7	0.059	0.068	-0.196	-0.149
x_8	0.056	0.060	-0.168	-0.125
x_9	0.040	0.076	0.299	0.101
x_{10}	0.037	0.073	0.294	0.073
x_{11}	0.040	0.073	0.302	0.079
x_{12}	0.040	0.074	0.311	0.105
x_{13}	0.074	-0.152	-0.119	0.367
x_{14}	0.072	-0.144	-0.072	0.355
x_{15}	0.061	-0.132	-0.149	0.366
x_{16}	0.074	-0.146	-0.091	0.344
x_{17}	0.066	-0.180	0.109	-0.306
x_{18}	0.062	-0.165	0.114	-0.294
x_{19}	0.063	-0.174	0.111	-0.294
x_{20}	0.062	-0.174	0.120	-0.293

$$-0.180x_{17} - 0.165x_{18} - 0.174x_{19} - 0.174x_{20}$$

$$\begin{aligned}
 f_3 = & -0.059x_1 - 0.085x_2 - 0.084x_3 - 0.095x_4 \\
 & -0.167x_5 - 0.186x_6 - 0.196x_7 - 0.168x_8 \\
 & + 0.299x_9 + 0.294x_{10} + 0.302x_{11} + 0.311x_{12} \\
 & - 0.119x_{13} - 0.072x_{14} - 0.149x_{15} - 0.091x_{16} \\
 & + 0.109x_{17} + 0.114x_{18} + 0.111x_{19} + 0.120x_{20}
 \end{aligned}$$

$$\begin{aligned}
 f_4 = & 0.017x_1 - 0.006x_2 - 0.028x_3 - 0.039x_4 \\
 & - 0.111x_5 - 0.135x_6 - 0.149x_7 - 0.125x_8 \\
 & + 0.101x_9 + 0.073x_{10} + 0.079x_{11} + 0.105x_{12} \\
 & + 0.367x_{13} + 0.355x_{14} + 0.366x_{15} + 0.344x_{16} \\
 & - 0.306x_{17} - 0.294x_{18} - 0.294x_{19} - 0.293x_{20}
 \end{aligned}$$

5 Conclusion

Through the above calculation and analysis, the following conclusions can be drawn:

- (1) There are 20 original evaluation indicators, but only 4 factors, which are far less than the number of evaluation indicators. Complex data are simplified and dimensionality reduced, and factors replace original variables to participate in data modeling, overcoming the defects caused by too many variables in the analysis process.
- (2) f_1 is the most important influence factor, with a contribution rate of 71.253%, which mainly reflects indicators such as “Standard operating procedure, Safety management system, Safety management organization, Safety management responsibility, Emergency equipment management, Emergency identification management, First aid materials management, First aid measures plan, Overall spatial layout, Security identifier setting, Clear evacuation route, Disposal of waste, Access control management system, Video monitoring system, Visual identification of hazards, Identification of unsafe behavior”.
- (3) f_2 is the “safety rules and regulations” factor, with a contribution rate of 14.806%. The establishment of effective safety rules and regulations is the basis of laboratory safety management, including four secondary indicators. Safety management system, including a series of provisions to ensure laboratory safety. Safety management responsibility, strengthen the responsibility of the post, clear the main responsibility of safety management and safety accidents should bear responsibility.
- (4) f_3 is the factor of “instrument and equipment management”, and the contribution rate reaches 6.622%. Instrument and equipment management mainly evaluates four aspects: general equipment management, special equipment management, protective equipment management and instrument safety management. It is necessary to build instrument and equipment ledger, and large equipment should have safe operating procedures and be on the wall.
- (5) f_4 is a “safety environment management” factor, with a contribution rate of 5.691%. Environment is an external and objective factor affecting laboratory safety. It includes the overall layout of the space, the setting of safety signs, the smooth evacuation channel, and the disposal of waste items. The overall layout of the space needs to be specifically designed according to the type and function of the laboratory, safety signs can remind the staff to prevent dangers so as to avoid accidents, emergency evacuation channel design takes into account the panic, fear and tension that may occur during the emergency evacuation process, waste items generally have inflammability, corrosion, toxicity and reactivity and other dangerous characteristics need special treatment. At the same time, it is necessary to reflect the different requirements of different laboratory types for the environment, so as to reflect the scientific and targeted evaluation system.

References

1. Ahmed, M., et al.: Multivariate statistical analysis of cosmetics due to potentially toxic/heavy metal(loid) contamination: source identification for sustainability and human health risk assessment. *Sustainability* **16**(14), 6127 (2024)

2. Souza, D.G.: HVO and biodiesel impact on diesel fuel stability: a multivariate data analysis approach. *Braz. J. Chem. Eng.* (prepublish): 1–17 (2024)
3. Forooghi, E., et al.: Detection of sheep butter adulteration with cow butter and margarine by employing Raman spectroscopy and multivariate data analysis. *Int. Dairy J.* **157**, 106010 (2024)
4. Gwashavanhu, K.B., Oberholster, J.A., Heyns, P.S.: A comparative study of principal component analysis and kernel principal component analysis for photogrammetric shape-based turbine blade damage analysis. *Eng. Struct.* **318**, 118712 (2024)
5. Goodluck, A., Otieno, J.D., Kosura, O.W.: Understanding farmers' perceptions on advisory services in Tanzania: comparative insights from principal component analysis and Q-methodology. *Heliyon* **10**(14), e34541–e34541 (2024)
6. McManus, B., et al.: Principal components analysis of driving simulator variables in novice drivers. *Transp. Res. Part F Psychol. Behav.* **105**, 257–266 (2024)
7. Chaouk, H., et al.: Application of principal component analysis for the elucidation of operational features for pervaporation desalination performance of PVA-based TFC membrane. *Processes* **12**(7), 1502 (2024)
8. Ongbali, O.S., et al.: Analysis of the key factors for small and medium-sized enterprises growth using principal component analysis. *Heliyon* **10**(13), e33573–e33573 (2024)
9. Maimoona, R., et al.: Laboratory safety climate assessment and its correlation with safety procedures amongst staff of a reference clinical laboratory. *J. Coll. Phys. Surg. Pak. JCPSP* **33**(11), 1259–1263 (2023)
10. Adams, H., et al.: Laboratory safety program fundamentals. *J. Am. Water Works Assoc.* **115**(7), 58–66 (2023)



Application of Computer Information Technology in Intelligent Analysis and Decision-Making Support of Diagnosis and Treatment Data

Yan Gao^(✉) and Yinsong Zhang

Yunnan Agricultural University, Yunnan, China
Y1856062879@163.com

Abstract. In view of the problem that TCM diagnosis and treatment data are complex, diverse and lack unified standards, this paper introduced computer information technology, combined with NLP and LSTM models, to improve the intelligent analysis ability of TCM diagnosis and treatment data, and provide scientific support for clinicians through an auxiliary decision-making system, thereby improving the efficiency and accuracy of diagnosis and treatment. First, this paper collected and integrated a large amount of medical records, treatment records, prescription information and other data from different TCM diagnosis and treatment platforms and medical institutions, standardizes data in different formats, and used natural language processing (NLP) technology for semantic analysis and data cleaning. Then, it built a classification and prediction model based on LSTM (Long Short Time Memory) to realize intelligent diagnosis and treatment recommendation generation for common diseases in view of the complex symptoms and individual differences unique to TCM diagnosis and treatment. Finally, based on the previous analysis and model results, an auxiliary decision system was developed to provide doctors with auxiliary decision suggestions in the treatment of complex symptoms and medication regimen recommendations during the diagnosis and treatment process. The results showed that after the use of the intelligent auxiliary decision system, the diagnosis and treatment time reduction rate was between 21.8% and 23.0%, and the accuracy rate reached 88.0%. Compared with traditional manual analysis methods, the efficiency of TCM diagnosis and treatment data processing has been significantly improved after the application of computer information technology. The introduction of computer information technology provides a new solution for the intelligent analysis and decision-making support of TCM diagnosis and treatment data, effectively solving the problems of complex, heterogeneous and insufficient standardization of TCM data.

Keywords: Computer Information Technology · TCM Diagnosis And Treatment Data · Intelligent Analysis · Decision-Making Support

1 Introduction

With the rapid development of big data and artificial intelligence, computer information technology has been widely used in various fields. Especially in the field of medical health, the accumulation of massive data has brought huge challenges to traditional analysis methods and also provided new opportunities for intelligent data processing. As a representative of traditional Chinese medicine, TCM has a history of thousands of years, and has accumulated a large number of medical records, prescriptions and treatment records in its diagnosis and treatment process. However, TCM diagnosis and treatment data has problems such as diverse formats, insufficient standardization, and complex symptoms, which limits its efficient use in the modern medical system. With the development of natural language processing (NLP) technology and deep learning algorithms, more and more studies have begun to apply these cutting-edge technologies to the analysis and mining of medical data. Through semantic analysis, data cleaning and other means, combined with the construction of a model based on the long short-term memory network (LSTM), the processing and analysis capabilities of TCM diagnosis and treatment data can be effectively improved.

This paper studies the collection, standardization, semantic analysis of TCM diagnosis and treatment data and its application in intelligent diagnosis and decision-making assistance. First, the paper integrated medical records, diagnosis and treatment records, and prescription information to build a comprehensive data set, and used NLP technology to clean and standardize the data to extract effective diagnosis and treatment information. Then, based on the LSTM model, the paper realized the intelligent diagnosis and treatment recommendation generation of common diseases for the complex symptom system of traditional Chinese medicine, aiming to develop a decision-making support system to provide scientific and intelligent support for clinicians in the treatment of complex symptoms and the recommendation of medication plans, and improve the efficiency and accuracy of diagnosis and treatment.

The structure of this paper is as follows: First, this paper introduces in detail the collection process of TCM diagnosis and treatment data and the data cleaning and standardization methods, especially how to perform semantic analysis and data processing through NLP technology. Then, this paper explores how to generate intelligent diagnosis and treatment suggestions based on the LSTM model, and analyzes the applicability and prediction accuracy of the model under different symptom systems. Finally, this paper demonstrates the development and application of the decision-making support system, discusses its performance in actual clinical scenarios, and how to improve the efficiency and accuracy of diagnosis and treatment through this system. Through the discussion of the above parts, this paper not only demonstrates the application potential of computer information technology in TCM diagnosis and treatment data processing, but also provides theoretical and technical support for the future development of intelligent TCM diagnosis and treatment. Such a structured introduction is not only closely integrated with existing research, but also clearly outlines the research purpose and methods of this paper, ensuring the innovation and practical value of the research.

2 Related Work

In recent years, with the rapid advancement of medical informatization, the application of computer information technology in medical data analysis has gradually become a research hotspot. Especially in the scenario of processing a large amount of unstructured data, the semantic analysis of medical records and treatment records using technologies such as natural language processing (NLP) has greatly improved the efficiency and accuracy of data processing. Dou et al. [1] constructed a TCM diagnosis and treatment data collection system for thyroid diseases based on real-world research based on computer information technology, providing a data collection platform and algorithm model for the evaluation of the efficacy of TCM treatment of thyroid diseases and pre-hospital intelligent triage. Peng et al. [2] summarized the diagnosis and treatment ideas and TCM drug compatibility rules for the clinical treatment of liver depression type erosive gastritis through data mining methods, and used computer information technology to provide a new method for the clinical experience research of the treatment of liver depression type erosive gastritis. Li et al. [3] believed that TCM, with its unique diagnosis and treatment techniques and clear curative effects, has made important contributions to serving the health of the whole people. However, in the face of the huge demand for health of the whole people, the service capacity and quality of the existing TCM diagnosis and treatment model have become increasingly obvious. The rapid development of the Internet and artificial intelligence technology has provided new opportunities for the upgrading and optimization of TCM diagnosis and treatment models. In the process of intelligentization of TCM diagnosis and treatment models, the mining and processing technology of TCM clinical diagnosis and treatment data is a key link in the intelligent computer information technology of TCM, and it has also become a bottleneck problem restricting the intelligentization of TCM. Yang et al. [5] believed that real-world TCM clinical evaluation is a method for comprehensively evaluating the clinical efficacy of TCM, aiming to deeply study the causal relationship between TCM intervention and clinical outcomes. Their method involves several key steps, including data integration and warehouse construction, and computer information technology. Li et al. [5] explored the clinical efficacy of comprehensive treatment of TCM model based on computer information technology on patients with pneumoconiosis through a pilot double-blind, randomized, placebo-controlled study. These technologies not only help doctors better understand patients' conditions, but also provide technical support for the construction of intelligent diagnostic systems. However, existing research is mostly focused on the field of Western medicine. Due to its unique symptom complexity and individual differences, TCM diagnosis and treatment data has not received sufficient attention and research.

At the same time, deep learning models, especially time series processing algorithms such as LSTM, have also achieved remarkable results in the application of medical data prediction and classification. This type of algorithm can track and predict the changes in patients' symptoms over a long period of time, providing the possibility for early detection and personalized treatment of diseases. Luo et al. [6] conducted an updated computer information technology analysis of traditional Chinese medicine compounds for the treatment of functional dyspepsia: a randomized, double-blind, placebo-controlled trial. Xie et al. [7] studied the diagnosis and treatment of rheumatoid arthritis with the

combination of traditional Chinese and Western medicine under computer information technology to improve efficacy and reduce toxicity. Wang et al. [8] studied the clinical efficacy of Chinese medicine in the treatment of grade 1 hypertension based on computer information technology and conducted a systematic review and meta-analysis. Jia et al. [9] studied the psychosomatic characteristics of the “Chinese medicine five elements person” in traditional Chinese medicine and summarized the data based on computer information technology. Ma et al. [10] reviewed the research on Chinese medicine treatment of premature ejaculation and explored the auxiliary efficacy of computer information technology. Xia et al. [11] used computer analysis to explore the comprehensive effect of Chinese medicine treatment on heart failure. Long et al. [12] studied a TCM diagnosis prototype for psoriasis based on computer information technology. However, most current studies still focus on structured numerical data analysis, lacking in-depth discussion of complex TCM diagnosis and treatment data. Existing models cannot provide sufficient support for the diverse symptom systems and treatment plans of TCM, and are difficult to adapt to the complexity of the TCM field.

3 Methods

3.1 Data Collection and Standardization

The diversity and heterogeneity of TCM diagnosis and treatment data are one of the main bottlenecks restricting its intelligent application. In order to solve this problem, first of all, this paper collects a large amount of medical records, diagnosis and treatment records, prescription information and other data from different TCM diagnosis and treatment platforms, hospitals and related medical institutions. These data come from a wide range of sources, covering patient groups in different regions and hospitals, and are highly representative.

3.1.1 Classification of Data Types

These raw data mainly include three categories: medical record data, diagnosis and treatment records and prescription data. Medical record data includes the patient’s personal information, description of the condition, and the doctor’s diagnosis; treatment records mainly include the time of consultation, symptom description, doctor’s observation and advice; prescription data covers the name of the prescription, dosage, and medication recommendations. Since most TCM data are unstructured texts, and the data formats of different platforms and institutions are inconsistent, this paper first formats the data in a unified manner.

3.1.2 Data Cleaning and Standardization

In order to ensure the accuracy and consistency of data, data cleaning is a key step. Using natural language processing (NLP) technology, the text data is first segmented to remove stop words, redundant data, and invalid information [13, 14]. For polysemous words and synonyms, this paper uses a word vector model to perform semantic disambiguation to ensure the semantic consistency of the text. In addition, in view of the differences in

the diagnosis and treatment data formats of different institutions, this paper standardizes them through custom rules and converts all data into the same format for subsequent analysis.

Symptom similarity calculation is used to measure the similarity between two symptom vectors and is often used in natural language processing and recommendation systems to recommend diagnosis or treatment plans based on the patient's symptoms. The calculation formula is as follows:

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| |\vec{B}|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

Among them, \vec{A} and \vec{B} represent the symptom vectors of two different patients, and the result range is $[-1, 1]$. The closer the value is to 1, the more similar the symptoms are.

3.2 Semantic Analysis and Feature Extraction

TCM diagnosis and treatment data is complex, mainly reflected in the diversity of symptom descriptions and the complexity of prescription structure. In order to extract effective information from unstructured data, this paper adopts NLP technology. First, the paper used the NLP method based on the pre-trained language model to perform semantic analysis on TCM medical records, and trained the model to understand the semantic expressions unique to TCM, such as the subtle differences between “cough” and “cough with phlegm”, so as to extract structured symptom features. Secondly, the paper extracted key feature information, including the patient's age, gender, main symptoms, duration of symptoms, past medical history, and the name, dosage, and usage of each medicinal material in the prescription. To ensure accuracy, an automated label generation system is designed to classify and annotate medical records. These features can be used as input for model training to help improve diagnostic accuracy.

3.3 Intelligent Diagnosis and Treatment Recommendation Generation Based on LSTM

After completing data cleaning and feature extraction, this paper constructs an intelligent diagnosis and treatment recommendation model based on LSTM. This model is good at processing time series data, learning the temporal relationship between inputs through medical records, symptoms and prescription information, and predicting future symptoms or diseases. A large amount of labeled data is used for training to ensure that the model accurately predicts different symptom combinations. The output is the probability distribution of disease occurrence, which assists doctors in making preliminary diagnoses. This model aims to improve diagnostic accuracy and generate effective treatment recommendations. In order to handle the different symptom severity of TCM diagnosis and treatment data, different weights can be assigned to different types of symptoms during model training. The weighted loss function formula is as follows:

$$L(y, \hat{y}) = \sum_{i=1}^n w_i \cdot (y_i - \hat{y}_i)^2 \quad (2)$$

Among them, w_i is the weight of the i -type symptom, y_i is the true value, and \hat{y}_i is the model's predicted value. This formula assigns different weights to different symptoms to ensure that the model pays more attention to more important symptoms.

Intelligent diagnosis and treatment suggestions: Based on the LSTM model, this paper further combines the expert knowledge base to generate treatment suggestions. The model first performs intelligent diagnosis based on the symptom characteristics in the medical record and gives possible disease types. Then, combined with the doctor's experience and the rule base in the expert system, the corresponding treatment recommendations are automatically generated, including recommended prescriptions and medication plans.

3.4 Development and Application of Decision-Making Support System

This paper develops a TCM decision-making support system, which includes data input, intelligent diagnosis and treatment recommendation modules. The data input module supports doctors to directly input or import medical records from the hospital system. The intelligent diagnosis module uses the LSTM model for real-time analysis and provides diagnosis results. The treatment recommendation module combines the expert knowledge base to generate medication and treatment plans, and doctors can adjust them based on experience, which enhances the flexibility and practicality of the system, aiming to improve the efficiency and accuracy of diagnosis and treatment. By analyzing the medical record data input by the doctor in real time, the system can quickly generate diagnosis results and treatment recommendations, greatly shortening the doctor's diagnosis time. The experimental results show that the system is highly accurate in dealing with common diseases, especially in cases where the symptoms are complex or the patient has a long medical history. The decision support provided by the system significantly reduces the workload of doctors. In addition, the system also shows high practicality in terms of medication recommendations, and can recommend personalized prescriptions based on individual differences of patients. The implementation process of the decision support system is shown in Fig. 1.

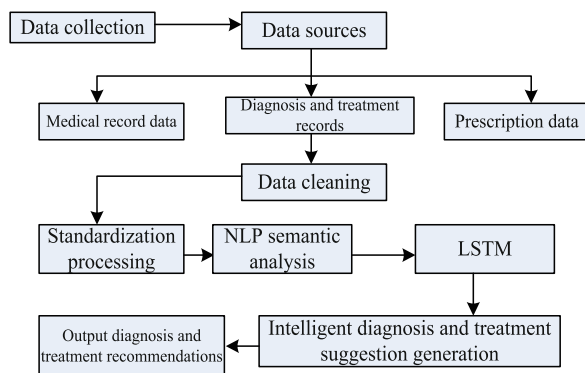


Fig. 1. Implementation process of the decision support system

Bayesian inference calculates the posterior probability of a disease through existing diagnosis and treatment data, and infers it based on the symptoms of new patients. The formula is as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (3)$$

Among them, $P(A|B)$ represents the probability of having disease A given symptom B , $P(B|A)$ is the probability of having symptom B under symptom A , and $P(A)$ is the prior probability of symptom A .

4 Results and Discussion

4.1 Experimental Environment and Parameter Settings

To verify the effectiveness of the TCM intelligent diagnosis and treatment system proposed in this paper, the experiment was conducted in an actual clinical environment, and the outpatient data of three TCM hospitals were selected as the experimental data source. The experimental platform is a Linux-based server equipped with an Intel Xeon processor, 256 GB memory, and an NVIDIA Tesla V100 graphics processor. TensorFlow and PyTorch were selected as the development framework, which are mainly used to build and train LSTM models. The NLP library in Python was used for data preprocessing and semantic analysis. The entire experimental system runs under GPU acceleration to ensure fast training and inference of the model. In order to better evaluate the effect of the model, the experimental data covers the medical records and diagnosis and treatment records of different patient groups, mainly including diagnosis and treatment data of common diseases such as respiratory diseases, digestive diseases, and rheumatism. The parameters of the LSTM model are set as follows: the time step is set to 10, the number of hidden layer units is 128, the optimizer uses Adam, the initial learning rate is 0.001, and the batch size is 64. In order to comprehensively evaluate the performance of the model, this paper sets the following key evaluation indicators: accuracy, recall, F1 value, and reduction rate of diagnosis and treatment time.

4.2 Result Analysis

(1) Intelligent diagnosis effect based on respiratory diseases.

In order to test the application effect of the system in common diseases, the first experiment selected the diagnosis and treatment data of respiratory diseases, including colds, coughs, bronchitis, etc. This experimental data set covers the medical records of 1200–1400 patients. The NLP and LSTM models predict the probability of disease occurrence and give diagnostic suggestions by learning the symptom changes in the patient's medical records. The results of intelligent diagnosis based on respiratory diseases are shown in Table 1.

Table 1 shows the data of four groups of patients in the intelligent diagnosis experiment of respiratory diseases, mainly including the number of patients, diagnostic accuracy, recall rate, F1 value and reduction rate of diagnosis and treatment time. Through the

Table 1. Results of intelligent diagnosis based on respiratory diseases

Patient Count	Accuracy (%)	Recall (%)	F1-Score (%)	Diagnosis Time Reduction (%)
1200	87.6	85.2	86.4	22
1300	87.8	85.4	86.6	22.5
1100	87.3	85	86.2	21.8
1400	88	85.5	86.8	23

analysis of the data, it can draw the following conclusions: First, the diagnostic accuracy of the system is between 87.3% and 88.0%, which shows that the system is relatively stable in different sizes of patient groups, and the accuracy rate increases slightly with the increase of the number of patients. In particular, in the experimental group of 1,400 patients, the accuracy rate reached 88.0%, which is a reflection of the strong adaptability of the system when processing more medical records, indicating that the performance of the model has been optimized under the expansion of scale. Secondly, the recall rate fluctuated slightly, from 85.0% to 85.5%. The recall rate reflects the system’s ability to identify positive examples. Although the recall rates in the four data sets are high, there is still slight room for improvement. The upward trend in the recall rate is consistent with the upward trend in the precision rate, which means that the system can maintain a high correct recognition rate when identifying common respiratory diseases. The F1 value, as the harmonic mean of the precision and recall rate, remains between 86.2% and 86.8%. The changes in the F1 values of each group are similar to the recall rate, and the F1 value increases with the increase in the number of patients, reflecting that the system performs well in balancing diagnostic accuracy and recall. Finally, the reduction rate of diagnosis and treatment time reflects the improvement of the system on the work efficiency of doctors. The reduction rate of diagnosis and treatment time ranges from 21.8% to 23.0%, indicating that after using the system, the average diagnosis and treatment time of doctors has been reduced by about 22%. As the number of patients increased, the time reduction rate also increased slightly, especially in the 1,400-patient group, where the treatment time reduction rate reached 23.0%, indicating that the system can more effectively improve the work efficiency of doctors under large-scale data sets.

(2) Personalized diagnosis and treatment of complex symptoms.

The experiment focused on the treatment of complex symptoms and selected rheumatism, a disease with diverse symptoms and large individual differences, for testing. The experimental data set contains long-term medical records of 1,000–1,300 rheumatism patients, with a data span of several years. The NLP and LSTM models not only predicted the progression of the patient’s condition, but also generated personalized treatment recommendations based on the medical records and expert knowledge base. The results of personalized diagnosis and treatment of complex symptoms are shown in Table 2.

(3) Medication suggestion generation effect.

Table 2. Results of personalized diagnosis and treatment of complex symptoms

Patient Count	Accuracy (%)	Recall (%)	F1-Score (%)	Diagnosis Time Reduction (%)
1000	81.3	78.9	80.1	18
1200	82	79.5	80.7	19
1100	80.8	78.4	79.9	17.5
1300	81.5	79.2	80.4	18.2

The experiment aims to evaluate the accuracy and practicality of the system in generating medication suggestions. The experimental data comes from the prescription information of 900 patients with digestive system diseases. The system recommends prescriptions based on the patient’s symptoms and diagnosis results, and adjusts the drug dosage. The effect of medication suggestion generation is shown in Fig. 2.

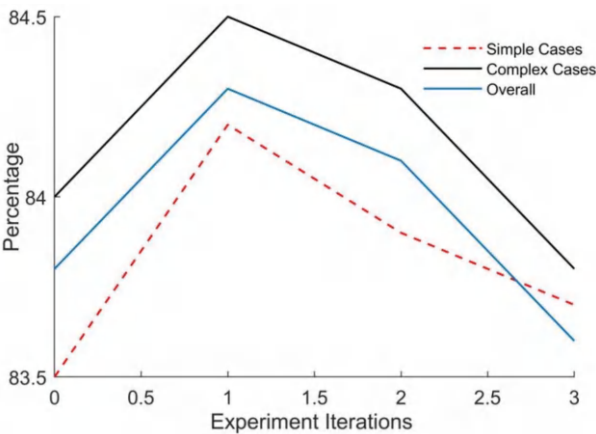


Fig. 2. Medication suggestion generation effect

Figure 2 shows the matching rate of medication suggestions for cases of different complexity of the intelligent diagnosis and treatment system. The matching rate of simple cases fluctuates slightly between 83.5% and 84.2%, with a maximum of 84.2%, indicating that the system is stable and accurate when dealing with cases with single symptoms and clear solutions. The matching rate of complex cases is slightly higher than that of simple cases, with a maximum of 84.5%, indicating that the system under computer technology can provide more accurate suggestions when dealing with complex cases in combination with LSTM model and expert knowledge base. However, as the number of iterations increases, the matching rate of complex cases gradually decreases, which may be related to the cumulative error. The diversity of symptoms and the complexity of prescription combinations increase the difficulty of matching. The overall matching rate is between the two, with the highest value being 84.3%. In multiple experiments, the system’s overall medication recommendation matching rate remained at a high level,

indicating that it has strong adaptability to diverse cases. However, similar to complex cases, the overall matching rate decreased in the later stage, which may suggest that there is a certain room for optimization in the system during long-term reasoning. In summary, although the system showed a high matching rate in the processing of both simple and complex cases, there is still room for improvement, especially in the stability of complex cases in multiple iterations. Future research can further optimize the model to maintain accuracy over a longer period of time, especially when dealing with complex diseases.

(4) Comprehensive performance of the system in clinical practice.

To further verify the comprehensive performance of the system in multi-disease scenarios, the last experiment selected the daily diagnosis and treatment environment of a traditional Chinese medicine clinic, covering patients with a variety of different diseases. The experiment lasted for one month and collected a total of 4,500 patients' diagnosis and treatment data.

The comprehensive performance of the system in clinical practice is shown in Fig. 3.

Figure 3 reveals the comprehensive performance indicators of the system in clinical applications, focusing on the changes in accuracy, recall, and F1 value with experimental iterations. The accuracy is stable, ranging from 84.4% to 84.7%, with small fluctuations, indicating that the system has strong adaptability and high diagnostic accuracy in multi-disease diagnosis. The recall rate was initially 82.8%, then rose to a peak of 83.0%, and slightly dropped to 82.8% in the fourth iteration, indicating that the system's ability to identify positive examples fluctuated slightly between different disease combinations, which may be related to the complexity of the disease and the quality of the data. The F1 value started at 83.6%, reached a maximum of 83.8%, and then fell slightly, reflecting the system's stable performance in balancing accuracy and recall. In summary, the system performed well in the multi-disease environment of the TCM clinic, especially in terms of accuracy. Future optimization should focus on improving the stability of the recall rate and reducing fluctuations to enhance the diagnostic reliability of the system in complex diseases. Through these measures, the system is expected to further improve its effectiveness in clinical decision support.

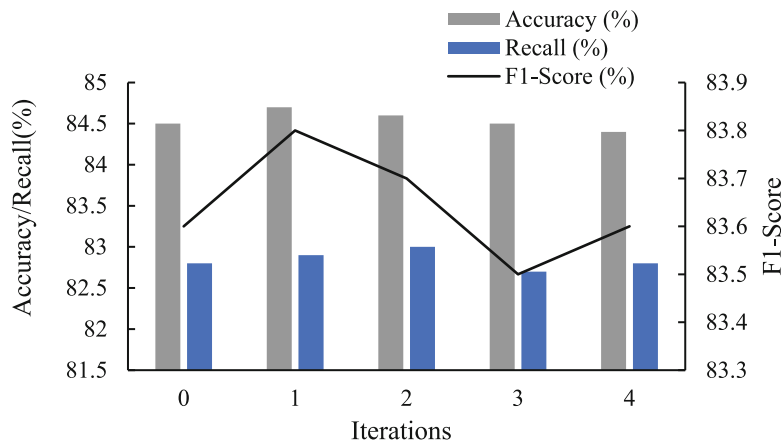


Fig. 3. Comprehensive performance of the system in clinical practice

5 Conclusions

This study proposed an innovative system based on NLP and LSTM models by applying computer information technology to the intelligent analysis and decision-making support of TCM diagnosis and treatment data. The main findings show that the system can effectively handle the heterogeneity and complexity of TCM diagnosis and treatment data, automatically generate diagnosis and treatment recommendations, and significantly improve the efficiency and accuracy of TCM diagnosis and treatment. In multiple experiments, the system’s performance in different disease scenarios has demonstrated its strong diagnostic capabilities and clinical applicability. In terms of practical application, the intelligent diagnosis and treatment system proposed in this study can not only provide effective auxiliary decision-making for doctors in TCM clinics, help simplify the diagnosis and treatment process of complex symptoms, but also significantly shorten the diagnosis and treatment time and reduce the workload of doctors. Its application potential is not limited to individualized treatment, but can also provide more technical support for TCM drug recommendations and medical big data analysis, and promote the modernization of TCM diagnosis and treatment. In general, this study provides an innovative solution for the intelligent processing of TCM diagnosis and treatment data, making up for the current limitations of TCM data standardization and insufficient analysis. The research results are not only of great significance in clinical practice, but also lay the foundation for in-depth research in the fields of medical informationization and personalized treatment in the future.

References

1. Zhili, D., Dongran, H., Yixing, L., et al.: Research on the construction of a TCM diagnosis and treatment data collection system for thyroid diseases based on real-world research. *Med. Soc.* **37**(6), 80–86 (2024)

2. Zhuoyin, P., Xiong, S., Renci, C., et al.: Analysis of TCM medication rules for erosive gastritis of liver depression type based on data mining. *World Latest Med. Inf. Abstr. (Electron. Ed.)* **021**(025), 1–4 (2021)
3. Xinlong, L., Peidong, H., Shuang, Z., et al.: Problems and challenges faced by intelligent mining of TCM clinical diagnosis and treatment data. *Chin. J. Tradit. Chin. Med.* **37**(12), 6962–6965 (2022)
4. Yang, W., Yi, D., Zhou, X.H., et al.: Translational analysis of data science and causal learning in real-world clinical evaluation of traditional Chinese medicine. *Sci. Tradit. Chin. Med.* **2**(1):57-65 (2024)
5. Li, J., Zhao, H., Xie, Y., et al.: Clinical efficacy of comprehensive therapy based on traditional Chinese medicine patterns on patients with pneumoconiosis: a pilot double-blind, randomized, and placebo-controlled study. *Front. Med.* **16**(5), 736–744 (2022). <https://doi.org/10.1007/s11684-021-0870-5>
6. Xiaoying, L., Yang, Y., Xinyong, M., et al.: Traditional Chinese medicine compounds for the treatment of functional dyspepsia: an updated meta-analysis of randomized, double-blind, placebo-controlled trials. *Digit. Chin. Med.* **4**(4), 273–289 (2021). <https://doi.org/10.1016/j.dcm.2021.12.003>
7. Xie, Z.J., Cao, W., Huang, L., et al.: Guideline for the diagnosis and treatment of rheumatoid arthritis with integrated traditional Chinese medicine and Western medicine to increase efficiency and reduce toxicity. *Tradit. Med. Res.* **8**(3), 15 (2023). <https://doi.org/10.53388/TMR20220805002>
8. Wang, J., Hua, Z., Li, C., et al.: The clinical efficacy of traditional Chinese medicine in the auxiliary treatment of grade 1 hypertension: a systematic review and meta-analysis. *TMR Modern Herb. Med.* **4**(4), 26 (2021). <https://doi.org/10.53388/MHM2021A1008002>
9. Jia, L.J., Yu, Y.X., Jun, Z.H., et al.: The psychosomatic traits of "people with the five elements in traditional Chinese medicine": a qualitative study. *Biomed. Environ. Sci.* **36**(11), 1068–1078 (2023)
10. Ma, D., Wang, A., Wang, H., et al.: A review of studies on the treatment of premature ejaculation with traditional chinese medicine. *Integr. Med. Nephrol. Androl.* (2024). <https://doi.org/10.1097/IMNA-D-24-00008>
11. Xia, L.L., Yang, S.Y., Xu, J.Y., et al.: Comprehensive effects of traditional Chinese medicine treatment on heart failure and changes in B-type natriuretic peptide levels: a meta-analysis. *World J. Clin. Cases* **12**(4), 766–776 (2024). <https://doi.org/10.12998/wjcc.v12.i4.766>
12. Long, H., et al.: A Prototype for diagnosis of psoriasis in traditional Chinese medicine. *Comput. Mater. Continua* **73**(3), 5197–5217 (2022). <https://doi.org/10.32604/cmc.2022.029365>
13. Deepa, K., Kumar, C.R., Rahman, M.K.G.E.D.: Mental health analysis using natural language processing. *J. Theor. Appl. Inf. Technol.* **101**(10):3688–3703 (2023)
14. Daming, L., Lianbing, D., Zhiming, C., et al.: Design of intelligent community security system based on visual tracking and large data natural language processing technology. *J. Intell. Fuzzy Syst.* **38**(6), 7107–7117 (2020). <https://doi.org/10.3233/JIFS-179789>



Security Vulnerability Detection and Defense of Smart Home Systems Based on the Internet of Things

Zhenghui Zhao and Miao Chen^(✉)

Chongqing Medical and Pharmaceutical College, Chongqing 401331, China
chenmiao@cqmpc.edu.cn

Abstract. With the widespread application of IoT (Internet of Things) technology in smart homes, the security protection of systems has received more attention. In response to the problems of incomplete vulnerability detection and outdated defense mechanisms in traditional methods, this article proposed an efficient and comprehensive vulnerability detection and defense framework. By periodically scanning all terminal devices in the smart home system through edge devices, known vulnerabilities in the system were identified. Through edge computing technology and AI (Artificial Intelligence) algorithm detection, lightweight distributed defense mechanisms were deployed between intelligent devices to ensure that when an attack occurred, it can quickly respond locally and reduce the system response time. The experimental results showed that all devices in the edge computing-based scheme had a high coverage rate of more than 90%. In terms of the average response time, the edge computing defense framework combined with AI algorithm was 7.2 s, far lower than the traditional defense methods. The experimental results prove that this research is efficient in improving vulnerability detection and defense response.

Keywords: Smart Home System · Internet of Things · Vulnerability Detection · Edge Computing Technology · Artificial Intelligence Algorithm

1 Introduction

With the rapid development of IoT technology, the application of smart home systems is becoming increasingly popular, but their security is also facing unprecedented challenges. The smart home system achieves remote monitoring and intelligent control of the home environment for users through the interconnection of IoT devices. This connection method brings many conveniences, but also makes the system vulnerable to network attacks and malicious intrusions. The current security protection mechanism mainly relies on centralized network security solutions, which are inadequate in dealing with diverse and decentralized smart home devices. Traditional vulnerability detection techniques mainly focus on a single device or local network system, making it difficult to fully cover complex smart home systems and facing the problem of insufficient vulnerability detection. The current defense systems mostly adopt passive response strategies

and lack the ability to actively identify and prevent new threats, which leads to the potential risks of security vulnerabilities being delayed in becoming apparent. Therefore, building an efficient and comprehensive security protection mechanism for smart home systems has become a key challenge that urgently needs to be addressed.

The research purpose of this article is to solve the key problems such as incomplete security vulnerability detection and lagging defense mechanism of smart home systems in the Internet of Things environment. Aiming at these problems, this article proposes a new vulnerability detection and defense framework that combines edge computing and AI algorithm. Through this framework, the shortcomings of traditional solutions can be solved. Through edge computing technology, distributed vulnerability detection can be deployed in the smart home system to ensure periodic scanning of all smart devices, timely discovery of known vulnerabilities, and full coverage of the system. This article utilizes AI algorithms for monitoring abnormal behavior and predicting attack patterns, actively identifying potential attack paths, and providing early warning and defense before an attack occurs, thereby enhancing the system's proactive defense capabilities. In addition, the defense framework processes security events locally through edge devices, reducing reliance on central servers, lowering defense system latency, and saving system resources, adapting to the resource constraints of IoT devices. This article proposes an innovative security framework to enhance the security of smart home systems, ensuring efficient and real-time detection and defense capabilities in the face of diverse attacks, while maximizing the optimization of device resource utilization.

2 Related Work

In the field of smart home security, various solutions have been proposed through research. In order to enhance the security and scalability of smart home systems, some researchers have proposed a new scalable blockchain architecture suitable for smart home systems composed of heterogeneous devices. By adopting efficient cross shard routing and sideblock schemes, throughput and latency can be improved, enhancing the scalability of the system [1]. Heshmati A proposed a blockchain-based authentication and access verification scheme that alleviates the security and efficiency challenges of smart homes by eliminating direct intermediaries and providing distribution [2]. Due to the high storage and computing costs of blockchain, it is difficult to promote in resource constrained devices. Other researchers have attempted to introduce defense strategies against signal attacks to enhance system security. In response to cross technology signal attacks, Zhang X designed a new real-time detection mechanism to distinguish between common ZigBee signals and analog signals, and improved passive defense strategies by misleading ZigBee signal eavesdropping [3]. Although these methods demonstrate high accuracy in attack detection, they lack strong adaptability to the heterogeneity of smart home systems. Therefore, there are still shortcomings in how to efficiently respond to diverse attacks and reduce device resource overhead.

In order to better solve the above problems, many researchers began to explore the combination of edge computing and artificial intelligence. Yang Y proposed a new method based on position sensitive hashing and time window technology to solve the problem of anomaly detection, achieving accurate and efficient detection [4]. However,

this method falls short in terms of defense response. Edge computing [5, 6] is a method that can reduce latency and lighten the burden of the central server. Some researches use edge devices to conduct distributed detection [7, 8], and use this detection method to monitor abnormal device behaviors in real time, which has improved the response speed of the system to a certain extent. However, the current edge computing scheme is still lacking in the comprehensiveness of vulnerability detection, because it cannot effectively respond to complex attack modes. Most existing defense mechanisms rely on historical data, and cannot predict unknown attacks. Facing these problems, this article proposes a vulnerability detection and distributed defense framework based on edge computing to solve the limitations of traditional methods in vulnerability detection and defense response.

3 Method

3.1 Smart Device Vulnerability Scanning

Periodic vulnerability scans are conducted on all terminal devices within the smart home system through edge devices [9, 10] to identify known vulnerabilities present in the system.

In the smart home system, for the security vulnerability detection of terminal devices, a periodic vulnerability scanning method based on edge computing is adopted, which combines parallel computing and feature matching technology to achieve an efficient and low resource occupation vulnerability identification process.

Set the scanning cycle for each device to T_i , where i represents the i th device in the system. The scanning interval is dynamically adjusted based on the type of device, workload, and historical risk factor. The scanning period T_i can be expressed as:

$$T_i = T_0 \cdot \frac{1}{\alpha \cdot R_i + \beta \cdot L_i} \quad (1)$$

T_0 is the initial set scanning time interval, and R_i represents the risk factor of the device, calculated based on the frequency and number of vulnerabilities that the device has been attacked in the past period of time. L_i represents the load factor of the device, reflecting the current usage of computing resources, while α and β are adjustment factors, balancing the impact of risk and load. In this way, by dynamically adjusting the scanning cycle, high-risk devices can be prioritized for scanning, and low load devices can also participate in detection in more frequent situations.

During vulnerability detection, edge devices utilize parallelization algorithms to simultaneously scan multiple terminal devices. Set the total scanning task as S , and divide the scanning tasks of each terminal device into S_1, S_2, \dots, S_n , where n is the total number of devices in the system. Using a parallel computing model, the scanning time T_s is calculated using the following formula:

$$T_s = \frac{T_{\text{Single device scanning}}}{p} + \frac{T_{\text{Synchronous overhead}}}{n} \quad (2)$$

$T_{\text{Single device scanning}}$ is the scanning time of a single device, p is the number of parallel processing threads, and $T_{\text{Synchronous overhead}}$ is the time for merging and synchronizing the scanning results of each device. This model can effectively reduce overall scanning time and improve scanning efficiency.

For identifying vulnerabilities, feature matching algorithms are used. Assuming that the feature vector in the known vulnerability library is $V = \{v_1, v_2, \dots, v_m\}$ and the detection data vector sent by the device is $D = \{d_1, d_2, \dots, d_k\}$, the conditions for vulnerability matching are:

$$M(D, V) = \sum_{i=1}^k \sum_{j=1}^m \delta(d_i, v_j) \quad (3)$$

$\delta(d_i, v_j)$ is the matching function. When the device data feature d_i matches the vulnerability feature v_j , $\delta(d_i, v_j) = 1$, otherwise it is 0. The matching function determines the existence of vulnerabilities by calculating the similarity between device data and known vulnerability features. When a vulnerability is matched, the edge device can immediately generate a vulnerability report and store the report information in the local database, triggering the corresponding defense mechanism. Through this periodic scanning mechanism, this article effectively improves the comprehensiveness and real-time performance of vulnerability detection in smart home systems, and ensures minimal resource utilization through reasonable task allocation and parallel computing.

3.2 Deployment of Distributed Defense Mechanisms

In the defense mechanism of smart home system, this article implements the distributed defense mechanism [11, 12] deployment through edge computing technology to ensure that the system can respond quickly when an attack occurs.

By embedding defense modules on each smart device, the module has a lightweight design that enables efficient operation on resource constrained devices. The AI algorithm used in the defense module is anomaly detection algorithm [13, 14], mainly based on behavior analysis and anomaly detection [15, 16]. Each device monitors its own network traffic, system calls, and device interaction patterns in real-time, and compares them with normal behavior models. When abnormal behavior is detected, defense mechanisms are immediately triggered. If the normal behavior feature vector is set as $B = \{b_1, b_2, \dots, b_n\}$ and the real-time observation data is set as $O = \{o_1, o_2, \dots, o_n\}$, then the abnormal judgment condition is:

$$A(O, B) = \sum_{i=1}^n |o_i - b_i| \quad (4)$$

When $A(O, B)$ exceeds the preset threshold, it is considered abnormal and the defense module is immediately activated.

To ensure real-time defense, the defense module collaborates with edge devices through low latency communication protocols. When a device detects a potential attack, it can immediately send an alert to the local edge device, and the edge device can coordinate with the defense modules of surrounding devices to enter a defense state

according to predefined defense strategies. This process avoids the delay problem caused by data transmission to the central server in traditional centralized solutions. The time from device detection to attack response includes the time of discovering anomalies, communication time between devices and edge nodes, and defense module response time.

The defense mechanism was designed using a distributed collaborative mode. Under the coordination of edge devices, various terminal devices in the system can share security status information and establish a collaborative defense network. When a device is attacked, edge devices can dynamically adjust the defense level of surrounding devices to enhance their protection capabilities.

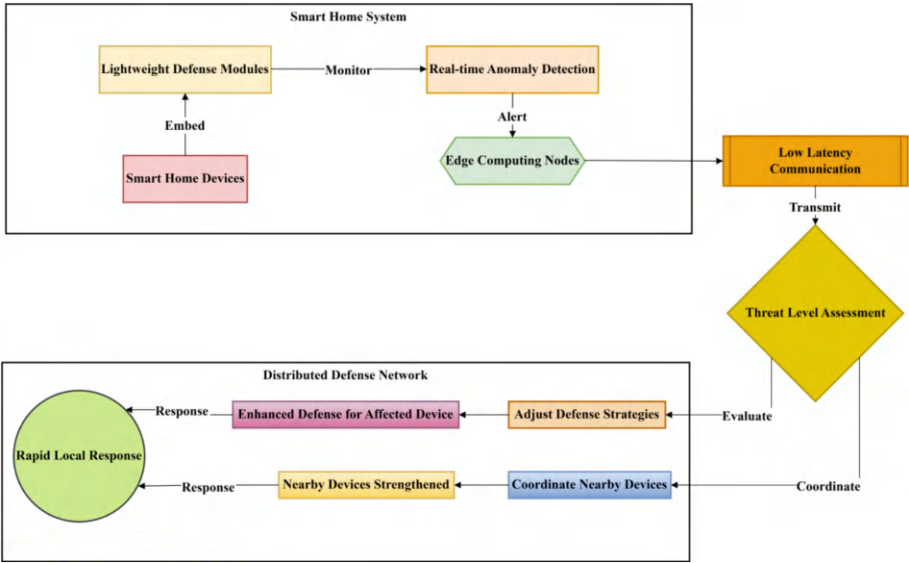


Fig. 1. System Distributed Defense Framework

Figure 1 shows the distributed defense framework of the smart home system, where each smart home device is deployed with a lightweight defense module. The module monitors the system through real-time detection technology and sends alerts upon detecting anomalies. The alarm is transmitted through the edge computing node, which ensures the timely transmission of information through low latency communication. Afterwards, the node can conduct a threat assessment and adjust defense strategies based on the severity of the threat, in order to effectively defend against the attacked device; Edge nodes also coordinate the defense strategies of neighboring devices, enhancing their protection capabilities by forming a distributed defense network. In the case of joint defense composed of multiple devices, the devices can prevent attacks from spreading in the network. The entire process ensures fast local response, utilizing localized distributed collaboration to prevent security incidents from causing larger scale damage and effectively reduce response time.

4 Results and Discussion

4.1 Vulnerability Detection Coverage Evaluation

In this chapter, the coverage of the vulnerability detection framework based on edge computing proposed in this article is evaluated through experiments. The detection coverage here is the ratio of the number of vulnerabilities detected by the system to the total number of vulnerabilities actually present in the system.

During the experiment, it is necessary to implant known vulnerabilities in the network environment for various types of smart home devices, and compare the proposed solution with traditional single device detection schemes for testing.

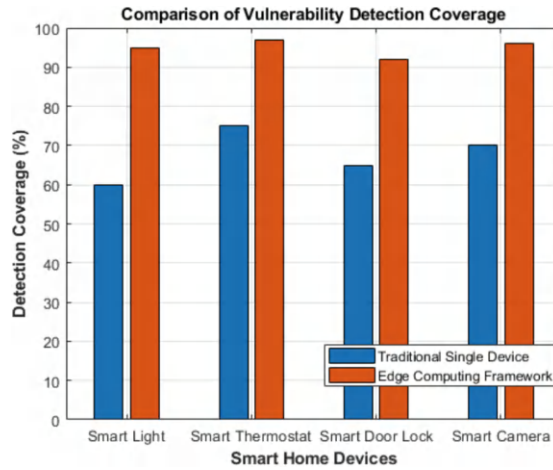


Fig. 2. Comparison of Vulnerability Detection Coverage

Figure 2 shows the comparison of vulnerability detection coverage between the traditional single device detection scheme and the detection architecture based on edge computing on four smart home devices: smart lights, smart thermostats, smart door locks and smart cameras. The horizontal axis represents the types of smart home devices, while the vertical axis reflects the percentage of vulnerability detection coverage. The coverage rates of traditional single device detection schemes on these four types of devices are 60%, 75%, 65%, and 70%, respectively. In contrast, the detection architecture based on edge computing has achieved 95%, 97%, 92% and 96% coverage respectively, and has a high coverage rate of more than 90% on all devices, which proves that the framework is effective in improving system security.

Table 1 shows the partial vulnerability detection results of two detection schemes. The framework based on edge computing has successfully identified three vulnerabilities; The traditional single device detection scheme failed to detect these vulnerabilities, demonstrating its limitations in detection.

Table 1. Analysis of Vulnerability Detection Cases

Vulnerability ID	Description	Detection Scheme	Detection Result(Yes/No)
V1	Device Authentication Vulnerability	Edge Computing Device	Yes
		Traditional Single Device	No
V2	Data Transmission Vulnerability	Edge Computing Device	Yes
		Traditional Single Device	No
V3	System Configuration Vulnerability	Edge Computing Device	Yes
		Traditional Single Device	No

4.2 Evaluation of Defense Response Timeliness

In this experiment, four different attack modes are used to compare the traditional defense methods (using firewalls or vulnerability detection tools) with the defense framework based on edge computing in this article.

Figure 3 shows four different attack types, namely DDoS (Distributed Denial of Service) attack, phishing attack, man-in-the-middle attack and ransomware attack. The response time of traditional defense methods and edge computing-based defense framework is compared with these four attack modes. The horizontal axis is the number of experiments, and the vertical axis is the response time.

In DDoS attacks, the response time of traditional defense methods fluctuates from 14 s to 17 s, showing a high delay overall, while the defense framework based on edge computing shows obvious advantages, and the response time is stable between 4 s and 6 s. This shows that edge computing can effectively and quickly respond to DDoS attacks, helping to reduce the time the system is affected. The sub graph of phishing attacks further proves this point. The response time of traditional methods is between 18 s and 22 s, and there is some volatility. The response time of edge computing-based solutions is between 5 s and 7 s, which shows its efficiency in responding to phishing attacks and can quickly identify and defend attacks. For man-in-the-middle attacks, the response time of traditional methods is generally 28 s to 32 s, while the response time of the defense framework based on edge computing is 9 s to 11 s. In the blackmail software attack graph, the response time of the traditional defense method is between 24 s and 27 s, while the response time of the defense framework based on edge computing is between 7 s and 9 s. Using these data, it can be clearly seen that the edge computing scheme has shown good response speed in various attacks, and the overall response time is between 4 s and 9 s, reflecting its indispensable role in smart home security protection.

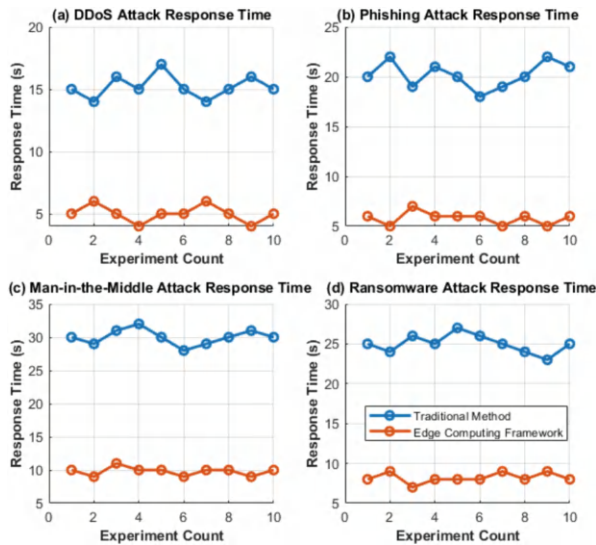


Fig. 3. Comparison of response time under different attack methods

Table 2. Average Response Time

Frame type	Attack Type (DDoS Attack)	Attack Type (Phishing Attack)	Attack Type (Man-in-Middle Attack)	Attack Type (Ransomware Attack)	Average Response Time(s)
Average Response Time (Traditional)(s)	15.2	20.2	30.0	25.0	22.6
Average Response Time (Edge Computing) (s)	5.0	5.8	9.8	8.2	7.2

Table 2 shows the average response time of traditional defense mechanisms and edge computing-based defense frameworks. The average response time of traditional defense methods is as high as 22.6 s, while the defense framework using edge computing combined with AI algorithm only takes 7.2 s, which shows good response ability.

5 Conclusions

This article proposes an innovative smart home system security vulnerability detection and defense architecture, which ingeniously combines edge computing and advanced artificial intelligence algorithms, and makes up for the limitations of traditional methods

in vulnerability identification and defense response. This research not only opens up new ideas for the security protection of smart home systems, but also provides an empirical basis, highlighting the key role of edge computing in enhancing security protection efficiency and response speed. The research in this article can be further deepened to explore how to optimize and upgrade the defense architecture in a more complex and ever-changing environment, in order to flexibly respond to continuously evolving network security threats and inject stronger impetus into the stable development of smart homes.

References

1. Liu, G., Wu, Z., Zhou, Y., et al.: Communitychain: toward a scalable blockchain in smart home. *IEEE Trans. Netw. Serv. Manage.* **20**(3), 2898–2911 (2023)
2. Heshmati, A., Bayat, M., Doostari, M.A., et al.: Blockchain based authentication and access verification scheme in smart home. *J. Ambient. Intell. Humaniz. Comput.* **14**(3), 2525–2547 (2023)
3. Zhang, X., Yu, S., Zhou, H., et al.: Signal emulation attack and defense for smart home iot. *IEEE Trans. Dependable Secure Comput.* **20**(3), 2040–2057 (2022)
4. Yang, Y., Ding, S., Liu, Y., et al.: Fast wireless sensor for anomaly detection based on data stream in an edge-computing-enabled smart greenhouse. *Digi. Comm. Netw.* **8**(4), 498–507 (2022)
5. Kong, X., Wu, Y., Wang, H., et al.: Edge computing for internet of everything: a survey. *IEEE Internet Things J.* **9**(23), 23472–23485 (2022)
6. Kong, L., Tan, J., Huang, J., et al.: Edge-computing-driven internet of things: A survey. *ACM Comput. Surv.* **55**(8), 1–41 (2022)
7. Xu, X., Tian, H., Zhang, X., et al.: DisCOV: distributed COVID-19 detection on X-ray images with edge-cloud collaboration. *IEEE Trans. Serv. Comput.* **15**(3), 1206–1219 (2022)
8. Cruz, P., Achir, N., Viana, A.C.: On the edge of the deployment: a survey on multi-access edge computing. *ACM Comput. Surv.* **55**(5), 1–34 (2022)
9. Laksmiati, D.: Vulnerability assessment with network-based scanner method for improving website security. *J. Comp. Netw. Architect. High Perform. Comp.* **5**(1), 38–45 (2023)
10. Darajat, E.Z., Sediyo, E., Sembiring, I.: Vulnerability assessment website E-Government dengan NIST SP 800–115 dan OWASP menggunakan web vulnerability scanner. *Jurnal Sistem Informasi Bisnis* **12**(1), 36–44 (2022)
11. Singh, A., Gupta, B.B.: Distributed denial-of-service (DDoS) attacks and defense mechanisms in various web-enabled computing platforms: issues, challenges, and future research directions. *Int. J. Semantic Web and Info. Sys. (IJSWIS)* **18**(1), 1–43 (2022)
12. Benmalek, M., Benrekia, M.A., Challal, Y.: Security of federated learning: Attacks, defensive mechanisms, and challenges. *Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle* **36**(1), 49–59 (2022)
13. Bahamid, A., Mohd, I.A.: A review on crowd analysis of evacuation and abnormality detection based on machine learning systems. *Neural Comput. Appl.* **34**(24), 21641–21655 (2022)
14. Friedrich, B., Sawabe, T., Hein, A.: Unsupervised statistical concept drift detection for behaviour abnormality detection. *Appl. Intell.* **53**(3), 2527–2537 (2023)
15. Selvathi, D., Chandralekha, R.: Fetal biometric based abnormality detection during prenatal development using deep learning techniques. *Multidimension. Syst. Signal Process.* **33**(1), 1–15 (2022)
16. Ibrahim, M.R., Youssef, S.M., Fathalla, K.M.: Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment. *J. Ambient. Intell. Humaniz. Comput.* **14**(5), 5665–5688 (2023)



In-Depth Discussion and Thorough Research on High-Availability Data Technology Within the Cloud Environment

Lei Yao^(✉)

The Esteemed College of Finance and Economics, Chengdu Polytechnic, Chengdu, Sichuan, China

18980882525@163.com

Abstract. Since the rapid growth of applications on the cloud for data, correspondingly, the number of databases, operation and maintenance requirements, and security requirements are rising simultaneously. There is an urgent need for high availability, scalability of various databases in the data cloud, and reducing Reduce operation and maintenance costs by means of automated operation and maintenance. This research encompasses key technical architectures like the scalability and high availability of heterogeneous databases and self-healing in case of failures, as well as key technologies for high availability of data. It provides a rapid recovery mechanism with the minimum number of slices to avoid data security issues caused by data damage or tampering. At the same time, it provides incremental merging of snapshot file systems to avoid data availability issues caused by excessive time consumption during the database recovery process.

Keywords: Database · Data Security · Cloud Data

1 Introduction

With the popularization and in-depth application of various data platforms of related companies, while enterprises possess a large amount of data, they also need to guarantee the storage security of data. At present, the data volume of large-scale systems is becoming increasingly huge, and the efficiency level in terms of data management and maintenance. is becoming more and more important. The need for fast, low-resource-consuming, and simple data management has become an urgent customer need to be addressed.

The inapplicability of traditional database architecture applications has triggered a series of problems in today's cloud environment. This inapplicability makes it difficult for database resources to fully utilize the powerful capabilities of cloud platforms in terms of flexibility, high reliability, and advanced self-healing services. In practical terms, traditional databases often cannot achieve an ideal operating state on the current virtualized cloud platform and cannot fully exert their performance advantages. This is like a high-performance sports car driving on a rugged road, making it difficult to show its true speed and passion, and thus has an adverse impact on business applications. Business

applications may experience sluggishness, delays, or even interruptions, bringing a bad experience to users and also hindering the operation and the advancement of enterprises [1, 2].

Also, the cloud environment itself has unique complexity. In this environment, data types are numerous and complex, and data forms are diverse. In the cloud environment, effectively backing up database resources is growing more and more challenging, and its complexity is constantly increasing. This is like finding the right path in a complex maze, full of challenges. If effective data backup in the cloud environment cannot be ensured, the cloud-based service of the database in the core business system will be seriously affected. Data may be lost or damaged, which is undoubtedly a huge loss for enterprises and may lead to the stagnation or even collapse of business [3, 4].

In light of the current pressing demand for the database utilization cloud services, this project emerges in response to the call of the times. Regarding various categories of database resources in the cloud setting, it proposes key technologies such as centralized monitoring, high availability, scalability, astute load balancing architectures, dynamic asset scheduling configurations, and high - efficient data protection systems. And algorithms. Through these technologies, functions such as installation and deployment, monitoring, fault handling, and the accomplishment of data protection for others databases in the cloud setting can be attained.. And in this process, standardized and automated application experiences and technical accumulations will likewise be furnished to supply robust technical backing for database cloud services and aid enterprises in evolving and progressing more favorably in the cloud epoch.

2 Related Works

Data resources in the cloud environment are distinct from present data resources[5, 6]. Traditional data backup means cannot truly safeguard data security.. The issue of backing up different kinds of data resources in the cloud environment needs to be resolved. It is necessary to attain the quick and operative backup of a variety of data resources in the environment [7, 8]. In order to ensure zero data loss, a rapid data backup and recovery solution needs to be provided in the cloud environment.

Databases based on physical resources are still deployed and delivered through manual intervention [9]. The cloud environment necessitates addressing the issues of automated and expeditious deployment and delivery of multiple databases. Standardized operation and maintenance services need to be furnished to enhance efficiency in operation and maintenance work [10].

In the cloud environment, there are multiple types of databases [11]. A unified high-availability and scalable architecture for various miscellaneous databases in the environment of cloud needs to be achieved [12]. The self - mending capacity of databases in the environment of cloud must be achieved, and the high scalability and availability of databases ought to be enhanced [13].

Databases in the cloud setting need to have automated intelligent resource scheduling and load balancing capabilities to achieve reasonable scheduling and use of server assets residing in the environment of cloud and solve the performance problems of various databases running in the cloud environment [14, 15].

3 Methods

3.1 In-Depth Investigation into the Structured Data Protection Strategy Based on Multi-dimensional Log Replication Technology

The objective is to attain super - efficient algorithms for securing structured data, actualize continuous data protection of databases, and guarantee zero data loss. The inquiry encompasses the following two aspects:

Exploration into the multi-dimensional log replication technology's structured data protection strategy.

On the basis of the multi-dimensional linked list IO log recording algorithm, by appending a TimeStamp and an occupant Unit Number to the data blocks one by one in the logical volume to signify the occupied storage unit and time, and establish an LMT (logic mapping table) which records the pointing position of the data block at the current time and an IRT (io recording table) that keeps a record of the size of the maximum backup set data. In accordance with the block index, this algorithm can rapidly regenerate data at any point in time; Formulate a retention time policy to automatically recover the space that the real-time protection data of the volume occupies [1].

Research on the technologies and strategies of structured database assurance.

Utilize RMAN for backup. RMAN represents a specific implementation of SMR (Server-Managed Recovery). It is an independent application furnished by Oracle. Through creating a client connection to the Oracle database, it is capable of accessing the internal duplicate and renewal data packages of the database.

The heart of RMAN is the command interpreter. The input commands are received by the command interpreter, and it converts these commands into remote procedure calls (RPC) to be executed on the database [5] (Fig. 1).

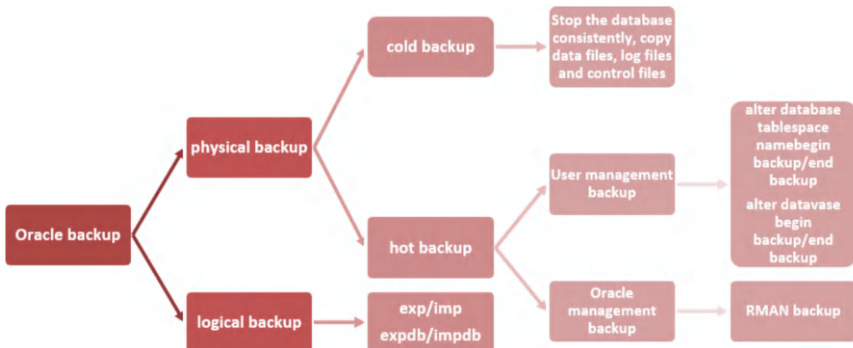


Fig. 1. Structured database protection technology

3.2 In-Depth Exploration on the Efficient Protection Tactics for Unstructured Data Founded on Various Alternative Algorithms

The client asynchronously distributes a copy of the write operation on the source volume to the mirror disk with the help of the volume filter driver (volume mirror driver). The

mirror volume is brought into being by the virtual block device driver (virtual block device driver) on the server side and provided for the client through the iSCSI protocol. Therefore, the server side can process and record IOs to realize the CDP backup function.

By adding a timestamp and an occupant unit number to each data block, with the Unit Number indicating the storage unit that is occupied, which one is it and the time indicating when it is occupied. Two types of tables are established respectively. One is the LMT (logic mapping table), and the other is the IRT (io recording table). The LMT records the position that the data block points to at the current time. The LMT always contains only the data of the current volume size. The IRT records when and by whom the data block is occupied. The maximum size of the IRT table is the size of.

3.3 Comprehensive Test and Verification of the Outcomes of High-Performance Data Protection Software

Based on our research results, a high-performance data protection software will be developed. The software will be tested and verified from the following aspects to confirm its ability of high-performance data protection:

In the context of the cloud platform of the data environment that needs protection, a disaster recovery center under the cloud platform can be correspondingly established for the entire big data environment. Once completed, the data of the production system that needs protection can be stored in the disaster recovery center for backup purposes in the cloud platform. If a disaster strikes the production center, the disaster recovery center of the cloud platform can quickly take over production business and continuously provide services. When the failure of the production center is remedied, the disaster recovery center will continue to offer services. Meanwhile, reverse replication is initiated to copy the data of the disaster recovery system back to the production system. After the initial data replication is accomplished and enters the real-time replication stage, the business can be switched back to the production center. Then the production center resumes providing services to the outside world, and then activates forward real-time replication. The new data of the production center is replicated to the disaster recovery center in real time. The research results also need to be tested and verified, collect user feedback, and further improve the software according to the feedback to ensure the promotion and application of the results in.

3.4 Technical Innovation Points

Through industry-leading technologies, it realizes rapid backup and recovery of a substantial amount of data and offers real-time preservation for structured and unstructured data. Conventional database backup technologies. With the most mature RMAN backup and renewal tool for Oracle databases, fail to satisfy the actual requirements in terms of the speed of data backup and recovery for large amounts of data. This research project investigates strategy for structured data protection using multi-dimensional log replication technology, the efficient protection strategy for researching unstructured data with diverse algorithms, the high-speed data regeneration technology founded on block tracking, the unstructured data efficient protection strategy based on multiple algorithms such as the multi-dimensional linked list IO log recording algorithm, the technology of

volume-level IO interception and separation, unstructured file backup, and the backup strategy for unstructured files which is based on multiple methods like the USN log. The main technical points are shown in the following Table 1. The implementation functions of each technical point are described later to achieve the purpose of ensuring the security and usability of numerous data systems in the cloudy scenario.

Table 1. Summary table of research technical points

	Technical point	Key technical principles
1	Instantaneous mounting technology for massive data at any point in time	Data copy management, storage virtualization technology
2	Intelligent automatic takeover technology	Real-time monitoring and analysis, intelligent decision-making and takeover trigger, seamless switching and recovery
3	Automatic verification technology for real-time protection data of volumes	Real-time monitoring and data acquisition, automatic verification algorithm, anomaly detection and alarm, data recovery and repair
4	Data and system rollback technology	Data backup and storage, system status recording, rollback trigger and decision-making, data and system restoration
5	Emergency takeover system	Real-time monitoring and early warning, rapid response and switching, fault diagnosis and repair, restoration and rollback
6	Automatic simulation disaster recovery drill technology	Scene simulation, system monitoring, automated execution, result evaluation

1) *Instantaneous Mounting Technology at Any time Point for Massive Data.*

The more extensive the backup data, the longer time required for restoration. Generally, when recovering a file, a typical data security system requires restoring all commerce data. This way is challenging to obtain the urgency guarantee result.

The subsystem of emergency take-over is founded on disk-level continuous data protection technology and a data block tracking and reorganization algorithm with high efficiency. It is capable of detecting, transmitting, and managing data modifications of application systems in sure time, rapidly backing up application data and the data within the operating system at any instant, and ensuring that the data can be made to revert to.

The subsystem of emergency take-over generates a imaginary disk from the standby version at any aforementioned time point, and mounts terabyte-level data and then presents it. Within minutes via the ISCSI/FC protocol to swiftly responsible for the application of the production server when the production server malfunctions.

2) *Intelligent automatic takeover technology*

Use artificial intelligence or machine learning methods to scientifically and accurately perform disaster recovery takeover. Intelligent takeover is more convenient

and flexible than automatic takeover. It does not require users to set parameters and thresholds [6]. It makes scientific and accurate judgments on whether to take over the user's production server according to different scenarios. (Fig. 2).

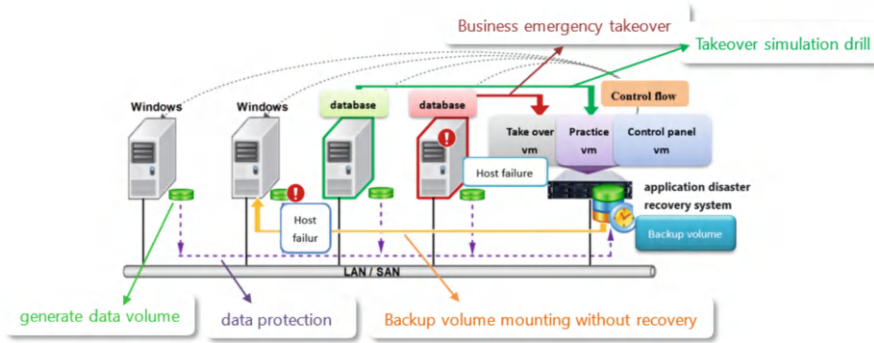


Fig. 2. Schematic diagram of intelligent automatic takeover technology

3) *Automatic Verification Technology for Real-time Volume Protection Data*

Automatically verify the consistency status [6], system logs, service status, database queries, etc. of the backup data generated by real-time volume protection through strategies, judge the integrity of the system disk view and data disk view data mapped over, ensure the availability of the backup data generated by real-time volume protection after recovery, discover problems in backup data in a timely manner, reduce the technical requirements and labor costs of existing manual drill verification, and increase the frequency of drill. This technology directly performs data verification in the shadow system of the production environment, with more comprehensive verification items and a higher degree of credibility [6].

4) *Data and System Return Technology*

When the production server failure is repaired, the data return function provided by the emergency takeover subsystem can start the system's own PE to write all data back to the production server, and then the production server continues to provide services outward [7].

5) *Emergency Takeover System*

When the emergency takeover system performs data return, in addition to returning the original data, it can also automatically write the newly generated data after takeover back to the production server, thereby saving recovery time and greatly reducing bandwidth usage [6].

The data return function of the emergency takeover system can support the return of business system data in various environments such as transitioning from a physical machine to a virtual machine, next from a virtual machine to a different virtual machine, and subsequently from a virtual machine to a physical machine to satisfy diverse user requirements to.

6) *Automatic Simulation Disaster Recovery Drill Technology*

Provide multi-level detailed definition functions for business, provide manual simulation drill function and automatic simulation drill function, support simulating the entire process from system startup to business use, and fundamentally ensure the availability of emergency measures and the availability and integrity of backup data. Provide simulation drills at any time point. Users can conduct takeover drills at any time point to verify whether the data at any time. The drill mode does not affect the normal operation of existing production businesses. After each automatic drill, the system will send a complete drill report to preset managers to facilitate managers to eliminate potential safety hazards in a timely manner [6].

This technology does not require. It can remarkably diminish the management pressure on operation and maintenance personnel, enabling users to acquire the up-to-date status of the production environment in a timely and efficient manner, eliminate potential safety risks, and save labor time costs. In this way, users can discover and eliminate possible safety hazards in a timely manner, thereby avoiding potential risks and losses. In addition, this technology can also greatly, promote more reasonable utilization of resources, and improve the efficiency and effectiveness of the entire production process. For both operation and maintenance personnel and users, this technology brings great.

4 Results and Discussion

4.1 Research Objectives

Conduct research on the current status and models of data protection models of various companies. Investigate the protection technologies and applications of structured and unstructured data with high availability in the core systems of companies. According to the application situation, study the high-availability data model.

Probe into the structured data defense strategy of multi - dimensional log duplication technology. Examine the technologies and strategies for structured database immunity. Obtain high-performance algorithms for structured data defense. Achieve continuous data protection of databases so as to warrant zero data loss.

Explore the efficient protection research strategy for unstructured data based on diverse algorithms. Analyze the efficient protection strategy for unstructured data based on multiple algorithms such as multi-dimensional linked list I/O log recording algorithm, volume-level I/O interception and separation technology, and unstructured file backup. Study the unstructured file backup strategy founded on multiple methods such as the USN log.

Our anticipated investigate objective is based on the cloud platform design schema. In combination with the technical characteristics of the cloud platform for computing, we explore the key technologies of high - efficiency data protection in the cloud

platform environment. Construct an independently controllable data backup system. Realize protection strategies and schemes for multiple types of data such as structured and unstructured data. And carry out test verification work for related functions. While ensuring the high timeliness of data backup, improve the quality of data backup. Provide technical experience accumulation for database protection in the cloud platform architecture environment. Provide better data protection functions for business system data. It is planned to be deployed on the cloud platform of the provincial power company. Carry out test verification work on high availability of cloud data of the database. Verify the effectiveness of the deployment scheme through on-site deployment. Conduct field test optimization. Further improve the deployment scheme to ensure the practical implementation of related applications of high availability software functions of cloud data of the database.

4.2 Testing and Verification of Achievements

We study the domestic and international application status and current application status of data backup products. Focusing on researching fast data regeneration technology based on block tracking. Conduct research on key technologies such as renewable technology of storage systems, real-time data incremental merging technology, data merging algorithm, and high-speed data recovery interface from the perspective of data security. Through high-speed data recovery interface and data on-site protection technology, build a data rapid recovery system. Address the issues of data recovery efficiency and excessive occupation of storage space. To insure the efficient operation of the company's storage system and provide a guarantee for data security in power grid informatization construction.

4.3 The Relevant Economic and Social Benefits

Enhance the enhanced availability of the database system to ensure the supreme availability of database services for business systems within the cloud environment. Presently, the operation of databases founded on physical resources in the cloudified milieu is executed as individual entities, incapable of attaining a comprehensive docking with the platform of cloud. The superior availability capability of the database in the cloudified environment cannot be effectively warranted. Through the implementation of this scheme, a highly scalable and available database cloud system architecture will be adopted. By integrating the technical traits of the cloud platform and leveraging its advantages, in the cloudified context, the superior availability of multiple databases will be realized, having no existence of a single - point - of - failure, averting the disruption of business systems due to database failures and lowering the operation and maintenance costs.

Ensure the protection of system data's security and lessen the economic losses occasioned by data destruction. Powerful data protection technology of high performance in the cloud environment investigated this time actualizes the functions of safeguarding structured and unstructured data, presents solutions for the duplicate and recovery demands of data files such as application system database data and distributed file systems, and lower the economic losses due to human-induced destruction or hardware impairment to various sorts of data.

Foster The promotion and application of domestic databases, heighten the capability of information security to be independently controllable, and stimulate the growth of national core competitiveness. Currently, the large-scale employment of foreign database systems has constantly posed potential threats to data security. As a result, when using an independent and traditional commercial database systems are replaced by a controllable database system, it is required to have the service capabilities of traditional commercial database systems. The database cloud system architecture with high availability and scalability developed in this project can be made compatible with mainstream domestic database systems and self-developed databases. The high availability is possessed by the cloud database constructed according to the research results of this project is further fortified, and it also has a robust self-healing ability in case of failures, shrink the performance disparity with conventional commercial database systems, encourage the diffusion and utilization of domestic databases, and thereby substantially improve national technological competitiveness.

5 Conclusion

With the continual progress of data cloud construction, a growing number of the deployment of business systems is carried out and run on the cloud, and the demand for setup of database cloud is becoming more conspicuous. The results of this research will be utilized in data cloud services. The data cloud will be endowed with the capacity to perform the deployment of various databases is agile and their operation and maintenance are automated, enhance the high scalability and availability of various cloud databases, be capable of dynamically scheduling and allocating database materials according to demand, facilitate on-demand dynamic allocation of resources in the cloud, and extend productive direction and functional support for the makeover of the database system.

Acknowledgements. This work was supported by the 2024 college-level scientific research project of Chengdu Polytechnic, “Research on highly reliable cloud storage data security technology based on blockchain technology (2024CZYG006)”.

References

1. Zhu, X., Qin, H., Wang, Z., Tong, F.: Research on Multi-Dimensional Linked List IO Emergency Takeover Log Recording Algorithm. Qinghai Electric Power (06) (2020)
2. Oracle+RMAN+11g+Backup and Recovery. <http://wenku.baidu.com>
3. An, G.: Data backup and recovery system for information innovation platform. In: Proceedings of the “Network Security Industry Development Forum” of the 2021 National Cybersecurity Publicity Week (10) (2021)
4. Wang, J., Huang, S.: Data backup and protection of hospital information system based on CDP technology. Digital Technology and Application (08) (2019)
5. Iovane, G., Di Gironimo, P., Benedetto, E., D’Alfonso, V.: Some Properties and Algorithms for Twin Primes. Applied Sciences
6. Naveen Kumar, C.G., Chandrasekar, C.: Advanced cryptography technique in certificateless environment using SDBAES. Int. J. Adv. Intel. Paradigms

7. Lee, H.-H., Stamp, M.: An agent-based privacy-enhancing model. *Information Management & Computer Security*
8. Wang, J., Kang, D.X., Zhang, A.J., Li, B.R.: Effects of psychological intervention on negative emotions and psychological resilience in breast cancer patients after radical mastectomy. *World journal of psychiatry*
9. Poolsappasit, N., Ray, I.: Towards Achieving Personalized Privacy for Location-Based Services. *Trans. Data Privacy*
10. Matteo, S., Massimo, G., Alfredo, C.C., Maura, M., Andrea, F.: Blooming in the rain. *Tumori*
11. Khayyat, M.M., Khayyat, M.M., Abdel-Khalek, S., Mansour Romany, F.: Blockchain enabled optimal Hopfield Chaotic Neural network based secure encryption technique for industrial internet of things environment. *Alexandria Engineering Journal*
12. Rajat, K.D., Pranam, P.: Block based Cryptographic Protocol with Arithmetic Operations. *Networking and Communication Engineering*
13. Hadeel Hadi, A.A., Rashed, Y.H.: Design of an alternative NTRU encryption with high secure and efficient. *International Journal of Mathematics and Computer Science*
14. Microsoft Technology Licensing LLC: Evolving Streaming Installation of Software Applications. In: Patent Application Approval Process (USPTO 20180060053). *Computer Business Week*
15. Xu, J., Hu, L., Sun, S., Xie, Y.: Cryptanalysis of countermeasures against multiple transmission attacks on NTRU. *IET Communications*



State Estimation and Fault Location of Multi-machine Power System Using Graph Neural Network and Variational Autoencoder

Fan Zhang¹(✉), Mengyan Guo², and Ya Wang¹

¹ School of Mechanical and Electrical Engineering, Wuhan Qingchuan University,
Wuhan 430204, Hubei, China
13986109556@163.com

² Wuhan University of Engineering Science, Wuhan 430204, Hubei, China

Abstract. With the continuous expansion and increasing complexity of electrical systems, traditional state estimation and fault location methods no longer meet the accuracy requirements of modern electrical systems. Graph Neural Networks (GNNs) and Variational Autoencoders (VAEs), as emerging machine learning technologies, have shown great potential in handling complex network structures and potential data representations. This article proposed a state estimation and fault localization method for a multi-machine power system that combines neural graphs and variational autoencoders. Firstly, neural graph networks were used to capture the topology and relationships between nodes in the energy system. Secondly, learning the latent representation of the system state through variational autoencoder can improve the accuracy of state estimation. Finally, combining the advantages of both, the fault can be quickly located. The experimental results indicate that the current system load is within the normal range of 110 nm to 140 nm. By monitoring the voltage level track, potential fault risks can be detected in a timely manner.

Keywords: Graph Neural Network · Variational Autoencoder · Power System · State Estimation · Fault Location

1 Introduction

In recent years, various scholars have conducted extensive research on state estimation and error localization of electrical systems. The data-driven approach uses machine learning algorithms to process large amounts of historical data and improve the accuracy of state estimation. The model-based approach focuses on constructing an accurate model of the electrical system to simulate the process of error propagation.

2 Related Works

In terms of improving the accuracy of state estimation in allocation systems, the autoencoder designed by Sundaray et al. is particularly outstanding, especially when facing situations that are difficult to directly observe [1]. Liao et al. studied many application

scenarios of neural graphs in electrical systems and emphasized their ability to handle complex data [2]. Khodayar et al. studied the latest advances in deep learning (DL) in power system research [3]. Maged et al. proposed a new time-varying high-dimensional process error detection framework that combined automatic change codes and short-term and long-term memory matrices [4]. Meanwhile, Ozcanli et al. also investigated the latest applications of DL in energy systems, including charge prediction and fault classification [5]. Jiang et al. innovatively proposed a temperature based graphical neural network model (TEMGNN), which achieved real-time status monitoring and fault warning functions for wind turbines [6]. Tong et al. successfully achieved accurate detection and classification of instantaneous transmission line errors using a neural network with a folded structure [7]. Tang et al. proposed an industrial process monitoring and fault isolation framework that combines an automatic variational encoder and a branch partitioning method [8]. To solve the problem of rotating machinery fault diagnosis, Feng et al. proposed a feature extraction method based on unsupervised neural graph network [9]. Finally, Zhang et al. proposed a semi-supervised method based on variational deep generative autoencoder model to solve the label scarcity problem in warehouse diagnosis and classification [10]. Although these methods have achieved some results, there are still some problems. For example, when it comes to state estimation and fault location of multi-machine power systems, data-driven methods may face problems of high computational complexity and insufficient generalization ability, and model-based methods are difficult to adapt to the rapid changes in the structure of the electrical system.

3 Method

3.1 Graph Neural Network Model

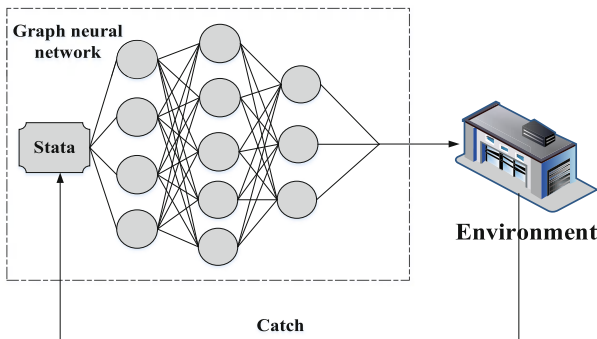


Fig. 1. Graph neural network model

Graph neural network model is a deep learning model that processes graph structured data. Its core lies in effectively learning the representation (also called embedding) of each node by aggregating information from the local neighborhood of the node. In the world of graph neural networks, entities are cleverly mapped to nodes, and the intricate

relationships between entities are intuitively presented through edges [11, 12]. This model not only captures the characteristic information of the nodes themselves, but more importantly, it also deeply explores the potential relationships between nodes connected by edges, thereby constructing a more comprehensive and accurate node representation. Figure 1 shows the structure of the graph neural network model.

The core idea of the model is to aggregate node neighbor information through a message passing mechanism and update the node feature representation. This process usually takes multiple rounds of iterations to capture information at a longer distance in the graph, and ultimately makes the feature representation of each node contain information about its neighbors and more distant nodes. It is explained through two key formulas. The first is the graph convolution operation formula of the graph convolutional network (GCN):

$$H(l+1) = \text{ReLU}(D_{12}AD_{12}H(l)W(l)) \quad (1)$$

$H(l)$ represents the node feature matrix of the l th layer, $W(l)$ represents the weight matrix of the l th layer, A is the result of adding the adjacency matrix to the self-loop, D is the degree matrix of A , and ReLU is the activation function. This formula implements normalized neighbor information aggregation and is used to update the feature representation of the node. The other is the attention weight calculation formula of the Graph Attention Network (GAT), which assigns different weights to the neighbors of each node:

$$\alpha_{ij} = \sum_{k \in (i)} \exp(\text{LeakyReLU}(a_{Whi \parallel Whj})) \quad (2)$$

Whi and Whj represent the feature vectors of node i and node j respectively, \parallel represents vector concatenation, α_{ij} is the attention weight between node i and node j , and K represents the set of neighbor nodes of node i . GAT introduces the attention mechanism to enable the model to dynamically adjust the way of information aggregation according to the importance of neighbor nodes.

3.2 Variational Autoencoder

Variational autoencoder is a deep learning model that combines the advantages of autoencoders and generative adversarial networks and can be used for different types of data generation and dimensionality reduction tasks. Its mathematical principle is mainly based on variational inference and Bayesian theory. The core idea is to approximate the data generation model through a variational distribution to achieve data encoding and decoding. VAEs consists of two parts: encoder and decoder. The encoder is responsible for mapping the input data to the latent space and learning the latent representation of the input data. The decoder generates data based on the points in the latent space, trying to reconstruct the input data and maximize the objective, the marginal logarithm of the input data. However, since directly optimizing this objective involves the integration of all possible latent representations, it is difficult. Therefore, VAEs use variational inference to approximate this problem and train the model by optimizing the sum of reconstruction loss and KL divergence. The reconstruction loss is the log-likelihood of the decoder

reconstructing the data, which is an expectation that the latent representation can be decoded back to a data point similar to the original data x . Therefore, the optimization problem of VAEs is usually solved by stochastic gradient descent and back-propagation algorithms. In this way, VAEs can learn a meaningful low-dimensional distribution to represent the data while keeping the reconstruction error low, and can generate new instances that are similar to the training data.

3.3 State Estimation and Fault Location System for Multi-machine Power System

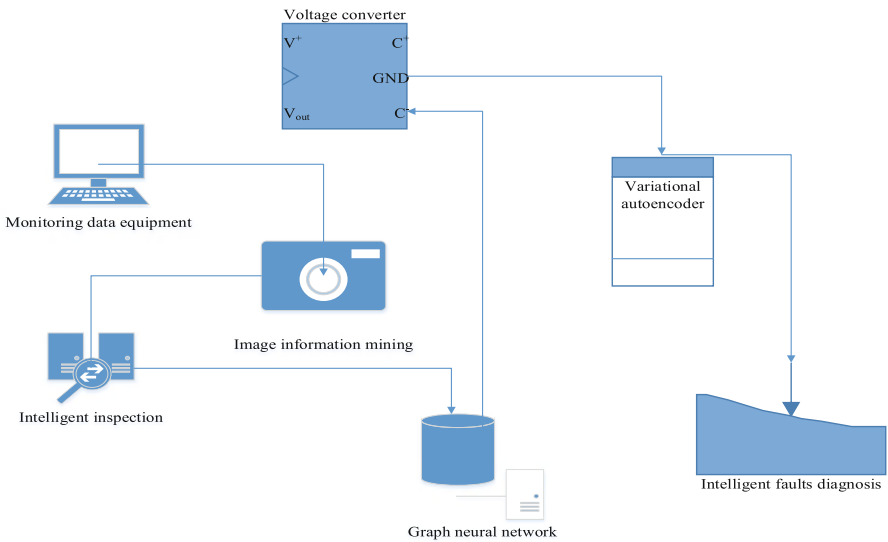


Fig. 2. Flowchart of fault positioning

In a multi-machine power system, in order to perform state estimation and fault location (the process is shown in Fig. 2), this article collects various operating data of the power system through monitoring data equipment, and uses image information mining technology to preprocess and extract features of the collected data [13, 14]. Then, the preprocessed data is input into the variational autoencoder. The variational autoencoder generates a model that can generate and reconstruct data by learning the latent representation of the data. Here, the variational autoencoder is used to extract the latent features of the power system state, which can reflect the normal operating state and possible failure modes of the power system [15]. At the same time, the structure of the power system is modeled using graph neural networks, which can capture the complex relationships between nodes. Here, the graph neural network represents the connection relationship and mutual influence between various devices in the power system, thereby achieving modeling of the overall state of the power system. After obtaining the potential features extracted by the variational autoencoder and the power

system structure modeled by the graph neural network, these two parts of information are combined for state estimation.

By comparing the difference between the actual monitoring data and the VAE reconstructed data, it is determined whether the current state of the current system deviates from the normal operating state. If there is a difference, it indicates a power grid fault. Finally, this article uses intelligent fault elimination technology to locate and diagnose potential faults in the current system based on the state estimation results and the potential features extracted from the VAE [16, 17]. Intelligent fault diagnosis technology can use neural network algorithms. It automatically identifies the type and location of the error based on the input data and characteristic information. These steps work together to obtain an accurate estimation of the power supply state and a fast location of the fault.

4 Results and Discussion

This article tests the performance of state estimation and fault location of a multi-machine power system, including load torque, stator current, and fault characteristic frequency.

4.1 Load Torque

Table 1. Load troque

Test number	Test conditions	Ambient temperature (°C)	Load troque (Nm)
1	Operating at rated voltage	25	120
2	80% rated voltage	30	135
3	Rated voltage, heavy load	22	110
4	90% rated voltage	28	140
5	Rated voltage, light load	20	125

This article examines the load torque value and ambient temperature of the fault monitoring system under different test conditions. These data play a key role in the state estimation and fault location of the subsequent performance operating conditions. In terms of the conditions mentioned above, Table 1 includes rated voltage operation, 90% rated voltage, rated voltage heavy load, 80% rated voltage and rated voltage light load. By recording the ambient temperature under various experimental conditions in the temperature range of 20 °C–30 °C, this article can see that the ambient temperature has an impact on the load torque and the overall state of the power system. By comparing the load torque values under different conditions, it can be said that the load of the drive system is within the normal range of 110 Nm to 140 Nm. By monitoring the voltage changes, potential error risks can be identified in a timely manner. This test data can be used as the data input required for training the model in this article. A graphical neural network with a variational autoencoder is used to estimate the state and location of faults

in the electrical system, build a suitable model structure, and use this data for training and optimization to achieve accurate estimation of the state of the electrical system and rapid fault location.

4.2 Stator Current

Figure 3 is calculated based on the sampling frequency and duration. Within this time range, the stator current signal is sampled and recorded. Stator current is a very important parameter in the power system, which directly reflects the operating status and load situation of the generator. By monitoring the changes in stator current, abnormal situations in the power system, such as overload, short circuit, imbalance, etc., can be detected in a timely manner. The frequency of the original stator current signal is 50 Hz, the amplitude is 3 A, and the phase offset is $\pi/4$. Most of the current values fluctuate between -8 A and $+8$ A. The frequency of the reconstructed signal of the stator current signal after being processed by the variational autoencoder is 120 Hz, the amplitude is 5 A, and the phase shift is $-\pi/6$. Its parameters are slightly different from the original signal to simulate errors or changes in the processing process. The current values of the reconstructed signal mostly fluctuate between -8 A and $+8$ A, but the waveform is not exactly the same as the original signal.

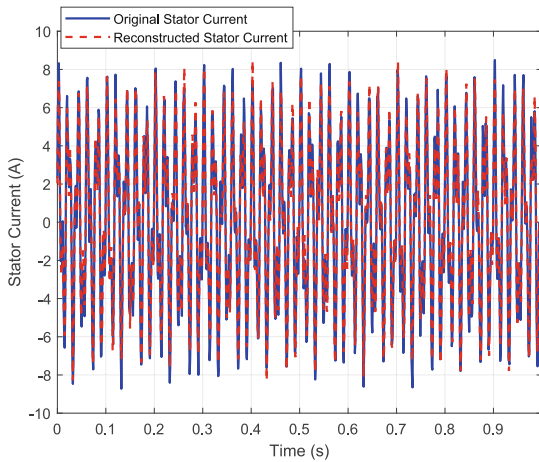


Fig. 3. Stator current

Due to slight differences in amplitude, phase offset and noise level, the reconstructed signal deviates from the original signal at certain time points. During the fault location process, the fault point or abnormal area can be identified through detailed analysis and comparison of the stator current waveform (such as the difference between the original signal and the reconstructed signal). If the stator current waveform of a generator suddenly changes significantly (such as an increase in amplitude, a change in frequency, etc.), it may mean that the generator is faulty or abnormal. By comparing the stator current waveforms of different generators, the propagation path and impact range of

the fault can be determined. Through careful analysis and comparison of the waveform diagrams, abnormal conditions in the power system can be discovered in a timely manner, providing strong guarantees for the safe and stable operation of the system.

4.3 Fault Characteristic Frequency

Table 2 clearly lists the characteristic frequencies of five different types of faults and their corresponding characteristic descriptions. For the short-circuit fault on busbar A, the harmonic frequency generated by the short-circuit current is 100 kHz; for the line-break fault on route B, the frequency generated by the traveling wave propagation is 50 kHz. The overload fault on transformer C caused a vibration frequency of 86 Hz; the ground fault on line D caused a frequency of 50 Hz caused by zero-sequence current; and the mechanical failure involving generator E caused a bearing fault with a frequency of 120 Hz. These data are presented in detail in a table format, which is of great reference value for fault diagnosis and location of power systems. They show the characteristic frequencies of different types of faults and their causes, which helps professionals to accurately analyze and handle faults.

Table 2. Fault feature frequency

Fault type	Fault location	Characteristic frequency description	Characteristic frequency values / kHz
Short circuit fault	Busbar A	Harmonic frequency generated by short-circuit current	100
Wire breakage fault	Route B	Frequency generated by traveling wave propagation	50
Overload fault	Transformer C	Vibration frequency caused by overload	86
Grounding fault	Route D	Frequency generated by zero sequence current	50
Mechanical failure	Generator E	Characteristic frequency of bearing failure	120

5 Conclusions

With the development of electrical systems, state estimation and fault detection have become key technologies to ensure their safe and stable operation. This article proposes a multi-machine electrical system state estimation and fault localization method that combines graph neural networks with variable autoencoders. This method fully utilizes the advantages of neural graphs in processing complex network structured data, as well

as the capabilities of variable autoencoders in data reduction and feature extraction, and can accurately estimate the state of electrical systems and locate fault points. The experimental results show that this method performs well in both state estimation and fault localization. Although this article has achieved some research results, there are still some gaps. (1) The complexity and computational efficiency of the model need to be further optimized; (2) Further research is needed to address the issue of estimating electrical system conditions and fault locations under extreme weather conditions. In response to the above shortcomings, future research may focus on the following aspects: firstly, further improving the computational efficiency and generalization ability of the model. By optimizing the model structure, improving algorithms, etc., it can reduce the complexity of the model and improve the calculation speed. Meanwhile, by introducing more diverse training data and testing scenarios, the model's generalization ability can be improved. The second is to study the application of this method in a wider range of electrical system scenarios.

References

1. Sundaray, P., Weng, Y.: Alternative auto-encoder for state estimation in distribution systems with unobservability. *IEEE Transactions on Smart Grid* **14**(3), 2262–2274 (2022)
2. Liao, W., Bak-Jensen, B., Pillai, J.R., et al.: A review of graph neural networks and their applications in power systems. *J. Modern Power Sys. Clean Ener.* **10**(2), 345–360 (2021)
3. Khodayar, M., Liu, G., Wang, J., et al.: Deep learning in power systems research: a review. *CSEE J. Power and Ener. Sys.* **7**(2), 209–220 (2020)
4. Maged, A., Lui, C.F., Haridy, S., et al.: Variational AutoEncoders-LSTM based fault detection of time-dependent high dimensional processes. *Int. J. Prod. Res.* **62**(4), 1092–1107 (2024)
5. Ozcanli, A.K., Yaprakdal, F., Baysal, M.: Deep learning methods and applications for electrical power systems: a comprehensive review. *Int. J. Energy Res.* **44**(9), 7136–7157 (2020)
6. Jiang, G., Li, W., Fan, W., et al.: TempGNN: a temperature-based graph neural network model for system-level monitoring of wind turbines with SCADA data. *IEEE Sens. J.* **22**(23), 22894–22907 (2022)
7. Tong, H., Qiu, R.C., Zhang, D., et al.: Detection and classification of transmission line transient faults based on graph convolutional neural network. *CSEE J. Power and Ener. Sys.* **7**(3), 456–471 (2021)
8. Tang, P., Peng, K., Jiao, R.: A process monitoring and fault isolation framework based on variational autoencoders and branch and bound method. *J. Franklin Inst.* **359**(2), 1667–1691 (2022)
9. Feng, J., Bao, S., Xu, X., et al.: Rotating machinery fault diagnosis based on feature extraction via an unsupervised graph neural network. *Appl. Intell.* **53**(18), 21211–21226 (2023)
10. Zhang, S., Ye, F., Wang, B., et al.: Semi-supervised bearing fault diagnosis and classification using variational autoencoder-based deep generative models. *IEEE Sens. J.* **21**(5), 6476–6486 (2020)
11. Remadna, I., Terrissa, L.S., Al Masry, Z., et al.: RUL prediction using a fusion of attention-based convolutional variational autoencoder and ensemble learning classifier. *IEEE Trans. Reliab.* **72**(1), 106–124 (2022)
12. Yuan, Y., Wang, Z., Wang, Y.: Learning latent interactions for event classification via graph neural networks and PMU data. *IEEE Trans. Power Syst.* **38**(1), 617–629 (2022)

13. Takiddin, A., Atat, R., Ismail, M., et al.: Generalized graph neural network-based detection of false data injection attacks in smart grids. *IEEE Trans. Emerg. Top. Computat. Intel.* **7**(3), 618–630 (2023)
14. Tama, B.A., Vania, M., Lee, S., et al.: Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. *Artif. Intel. Rev.* **56**(5), 4667–4709 (2023)
15. Luo, G., Cheng, M., Hei, J., et al.: Stacked denoising autoencoder based fault location in voltage source converters-high voltage direct current. *IET Gener. Transm. Distrib.* **15**(9), 1474–1485 (2021)
16. Ismail, M., Takiddin, A., Atat, R., et al.: Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids. *IEEE Trans. Artif. Intel.* **5**(3), 1287–1301 (2023)
17. Yu, J., Zhang, Y.: Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review. *Neural Comput. Appl.* **35**(1), 211–252 (2023)



Fault Detection and Diagnosis of Ship Circuit Based on Machine Learning Algorithm

Shuyan Liu^(✉)

Shanghai Maritime University, Shanghai 201306, China
syliu@shmtu.edu.cn

Abstract. With the continuous progress of ship technology, the detection and diagnosis of circuit faults has become particularly important, but traditional methods often rely on manual experience, resulting in long fault response time and low diagnostic accuracy. The purpose of this paper is to use machine learning algorithms to improve the efficiency and accuracy of ship circuit fault detection and diagnosis. Firstly, this study collects historical fault data of ship circuits, including sensor readings, voltage, current, etc., and performs data preprocessing to remove noise and missing values. Subsequently, this study employs feature selection methods to extract key features and utilizes various machine learning models (random forest and support vector machine) for training and validation. The experimental results show that the accuracy rate of the random forest model on the test set has reached 92%, while the accuracy rate of the support vector machine is 88%. The conclusion shows that the fault detection and diagnosis method based on machine learning can significantly improve the speed and accuracy of identifying ship circuit faults, and provide effective technical support for the safe operation of ships.

Keywords: Random Forest · Support Vector Machine · Ship Circuit Fault Detection · Noise Removal

1 Introduction

With the rapid development of the global shipping industry and the continuous progress of ship technology, the complexity of circuit systems has also increased. Circuit failure will not only affect the safety and reliability of the ship, but may also lead to serious economic losses and environmental pollution. Therefore, timely and effective detection and diagnosis of circuit faults has become an important issue in ship operation and management. However, traditional fault detection methods often rely on manual experience, resulting in long fault response times and low diagnostic accuracy, which is particularly unsuitable in the rapidly changing shipping environment.

In order to solve this problem, this paper aims to use machine learning algorithms to improve the efficiency and accuracy of ship circuit fault detection and diagnosis. By collecting historical fault data of ship circuits and performing data preprocessing to remove noise and missing values, this paper adopts feature selection method to extract

key features, and uses various machine learning models such as random forest and support vector machine for training and verification. This method not only improves the speed of fault detection, but also improves the accuracy, and provides effective technical support for the safe operation of the ship.

The structure of the article is arranged as follows: Firstly, the article introduces the relevant background and research significance, followed by a detailed description of the data collection and processing process, feature selection, and specific methods of model training. Then, the article presents experimental results and analyzes them, and finally summarizes the research conclusions and looks forward to future development directions. Through the research of this system, it is expected to provide new ideas and practical basis for the detection and diagnosis of ship circuit faults.

2 Related Work

The technology of fault detection and diagnosis of ship circuits is constantly developing, and related research focuses on improving the accuracy and response speed of fault identification. A number of scholars have proposed innovative methods for different types of failures and technical methods, which have provided important support for the safe operation of ships. Huo Yanfei proposed a ship simulation circuit fault diagnosis method based on RBF neural network to address the problem of complex interaction of ship simulation circuit components, difficulty in highlighting fault signals in a large number of normal signals, and difficulty in extracting and identifying fault features [1]. Wang Zhuofan discussed the causes of ship arc faults in response to the increasing number of ship fire accidents caused by electric propulsion ship arc faults, and analyzed the time-domain and frequency-domain characteristics of arc faults in AC/DC and series parallel circuits [2]. Liu Ruijuan designed an accurate acquisition system for low-power electronic circuit faults in ship engines to improve the detection efficiency. The simulation test results showed that the low-power electronic circuit fault detection accuracy of the ship's main engine is high, reducing the error of low-power electronic circuit fault detection of the ship's main engine [3]. Liu Yafei used an SVM based ship electronic fault classification model, taking the collected electronic equipment circuit operation signals as classification samples, classifying and identifying electronic fault states, and storing the classification results in the database unit database in the form of ER tables by the cataloging unit [4]. Yang Shaolong realistically presented the operation scene of the ship's piping system, the composition of the electrical control cabinet and electrical components, with a focus on reproducing the real fault circuit state and troubleshooting process, achieving random simulation of faults and autonomous measurement of voltage and resistance at any node [5]. Tsaganos G evaluated intelligent diagnostic methods applicable to two-stroke slow speed marine diesel engines, with the aim of promoting effective detection and classification of faults that occur [6]. Ellessen A L proposed a spectrum anomaly detection algorithm independent of fault types for detecting degradation of marine diesel engines in autonomous ferries [7]. Cheliotis M aimed to develop a new ship diagnostic framework based on operational data and fault probability, thereby enriching relevant literature [8]. Orhan M conducted a systematic review of fault detection and diagnosis models specifically designed for marine machinery and systems. Through a comprehensive review of literature from 2002 to 2022, the

number of 72 core articles was highlighted [9]. Jiang Y monitored and analyzed the engine status in real-time through FPGA image scanning, and preprocessed the engine operating status data using step tracking technology to make it a standard signal [10]. Various advanced fault detection methods and models provide guarantees for the safety and reliability of ship circuits.

3 Method

3.1 Source of Historical Fault Data

Various sensors are installed on the ship to monitor the voltage, current, temperature and other key indicators of the circuit in real time. The data generated by these sensors will be centrally stored in the ship’s monitoring system. In addition, the ship will undergo regular maintenance and inspections in its daily operations, and related fault reports and maintenance records will also be included in the scope of historical data. These documents usually list in detail the time, place, type of failure and its repair measures of the failure, which provides valuable information for subsequent data analysis. The ship’s operation management system may also record event logs related to circuit failures, including alarm information and operator response time [11]. The comprehensive use of these data can not only help identify potential failure modes, but also provide rich samples for the training of machine learning models, and improve the accuracy and efficiency of fault detection and diagnosis. By analyzing these historical data, the key characteristics that affect circuit failures can be effectively extracted, thus providing a solid foundation for subsequent fault detection methods. Table 1 is a copy of fault data collected and collated:

Table 1. Fault data

Fault ID	Fault Occurrence Time	Fault Type	Voltage (V)	Current (A)	Maintenance Action	Fault Status
001	2023-01-15 10:30	Short Circuit	24	15	Replaced damaged cable	Repaired
002	2023-02-20 14:45	Overload	12	25	Adjusted load	Repaired
003	2023-03-05 09:15	Open Circuit	0	0	Replaced fuse	Repaired
004	2023-03-18 11:00	Circuit Fault	10	20	Checked and repaired wiring	Repaired
005	2023-04-12 16:30	Short Circuit	24	18	Replaced damaged component	Repaired

3.2 Noise Removal and Missing Value Processing

Noise usually refers to information that introduces misleading or random fluctuations in the data set, which may be due to sensor instability, external interference, or system failure. During processing, moving average or median filtering is used to reduce random fluctuations in the data while retaining important trends and characteristics. For the processing of missing values, K-neighbor (KNN) padding and Multiple Interpolation are used [12]. These methods can effectively speculate on missing data, thereby reducing the deviation and information loss caused by missing values. In practical applications, through effective processing, the overall quality of the data set can be improved, the training effect and generalization ability of the machine learning model can be enhanced, so as to provide more accurate prediction and analysis in fault detection and diagnosis.

3.3 Feature Selection and Extraction

Supposing we are analyzing the circuit system of a cargo ship and collecting real-time data obtained from a variety of sensors, including voltage, current, temperature, humidity and corresponding fault history records. First, the data preprocessing stage will clean the original data to remove noise and missing values. Then, through statistical analysis and domain knowledge, the key characteristics related to circuit failures are identified [13]. In the case of circuit overload, the current reading is usually significantly higher than the normal range; in a short-circuit fault, the voltage will drop abruptly. These characteristics can be extracted by calculating statistics such as their mean, variance, maximum, and minimum values. Using time series analysis technology, trends and cyclical characteristics can be extracted from historical data, which are particularly important for fault prediction.

In the feature selection process, the feature selection algorithm recursive feature elimination (RFE) and the importance evaluation method of random forest are used to filter out the most distinguishing features. Through RFE, we can gradually eliminate the features with less impact, and finally select a set of feature sets with the best classification effect. At the same time, the use of random forest model to calculate the importance score of features can help identify the most predictive features under different fault types.

3.4 Machine Learning Options

In this paper, random forest and support vector machine (SVM) are selected. Each of these two models has unique advantages and is suitable for processing complex fault data. Random forest is an integrated learning method that improves prediction performance by constructing multiple decision trees and combining their outputs. Its main advantage lies in its strong robustness to noise and overfitting. Random forests can construct each tree by randomly selecting subsets of data and subsets of features, thereby introducing diversity in the training process [14]. This characteristic makes random forests particularly suitable for dealing with high-dimensional data and feature selection problems. In circuit fault detection, random forest can effectively identify various fault modes and help analyze which electrical parameters are most critical to fault prediction through characteristic importance assessment. In addition, due to its built-in cross-verification mechanism,

random forest performs well when processing unbalanced data, which can improve the recognition rate of a few types of failures. Support vector machine (SVM) is a powerful classification algorithm, especially suitable for small samples and high-dimensional data sets. SVM realizes classification by finding the optimal hyperplane in the feature space to maximize the interval between classes. In ship circuit fault detection, the advantage of SVM lies in its ability to handle complex decision boundaries and can effectively distinguish between different types of fault states [15]. By using kernel functions, SVM can find the optimal classification surface in the nonlinear feature space, thereby adapting to the nonlinear characteristics of the data. In addition, SVM performs well on high-dimensional data, which is suitable for processing multi-dimensional sensor data in ship circuits.

4 Results and Discussion

4.1 Experimental Setup

Designing the experimental setup and environments is meant to support accurately fault detection and diagnosis of ship circuits by machine learning algorithms. This method will rely initially on the conducting of the experiment in a laboratory setting provided with high-performance computing resources using a multi-core CPU and a server with large-capacity bodily memory to satisfy large-scale data processing and complex model training needs. Software-wise, Python will be the major programming language due to its range of machine learning libraries, including Scikit-learn, TensorFlow, and Pandas, which are used for data processing, feature extraction, model training, etc. Moreover, to address repeatability and reliability of results in the experiment, Jupyter Notebook will be opted for experiment logging and data visualization.

The experiment will be attached to the ship's live monitoring system for acquisition of circuit-related sensor data which include voltage, current, temperature, etc. during normal and faulty operational modes. Data are stored based on SQL database management systems that facilitate efficient query execution and data management. Generally, the stage of the experiment is composed of 4 parts which include data preprocessing, feature selection, model training, and model test. Data preprocessing involves the use of NumPy and SciPy to prepare the data for analysis. The random forest importance assessment method will be used in the feature selection process to identify important features. The model training is carried out based on different algorithms including random forests and support vector machines maximizing the performance using cross-validation on exhaustive hyper-parameter tuning. Then, the two algorithms are evaluated against each other on detection accuracy and diagnosis time for justifying their benefits in ship fault detection and diagnosis.

4.2 Experimental Results

The two algorithms were subjected to 20 repeated experiments to test the performance of the diagnostic accuracy, as shown in Fig. 1:

After 20 repeated experiments on the diagnostic accuracy of random forest (RF) and support vector machine (SVM) in ship fault detection, the performance of the two

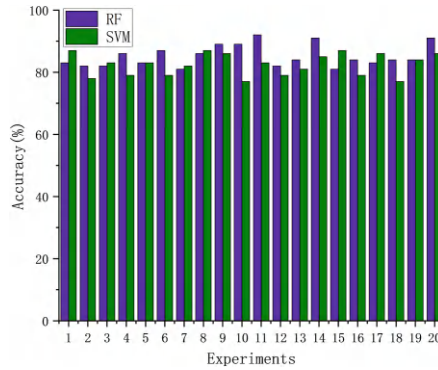


Fig. 1. Detection accuracy

algorithms can be extracted from the data. First, the accuracy data of random forests ranges from 81% to 92%, while the accuracy of support vector machines ranges from 77% to 88%. The average accuracy rate of random forests is 84.5%, while the average accuracy rate of support vector machines is 82.4%. Although the performance of the two algorithms is close, random forest has shown higher accuracy in multiple experiments, especially reaching the highest value of 92% in the 11th experiment, indicating its advantages in complex failure modes. Judging from the standard deviation of the data, the accuracy rate of random forests fluctuates relatively little, showing its stability, while support vector machines show large fluctuations in some experiments, especially in the 10th experiment. The accuracy rate is only 77%, which may be affected by data characteristics or model parameter settings. Figure 2 shows the test results of the diagnosis time:

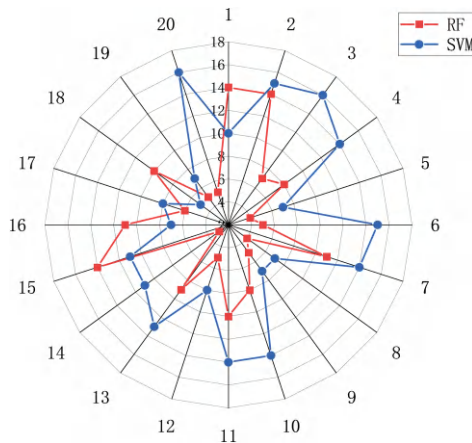


Fig. 2. Diagnosis time (s)

The data analysis of the ship's fault diagnosis time (seconds), the test results of random forest (RF) and support vector machine (SVM) showed significant differences. First, the diagnosis time of random forest ranges from 3 s to 14 s, while the diagnosis time of support vector machine is between 5 s and 16 s. The average diagnosis time of random forest is 8.1 s, which is significantly lower than the average diagnosis time of support vector machine, which is 11.2 s. This result shows that random forests are more efficient in the process of fault diagnosis. In terms of data volatility, the diagnosis time of random forest is relatively stable, and the time is short in multiple experiments, which shows its ability to respond quickly when dealing with fault diagnosis tasks. In some experiments, such as experiment 3 and experiment 2, the diagnosis time of support vector machines is as high as 16 s and 15 s, showing large time fluctuations, which may be related to the complexity of the model when dealing with specific fault characteristics.

5 Conclusion

Based on machine learning algorithms, especially random forests and support vector machines, this paper discusses the effectiveness of ship circuit fault detection and diagnosis. The experimental results show that the random forest is superior to the support vector machine in terms of accuracy and diagnosis time, and shows its advantages and stability in complex failure modes. This research provides effective technical support for real-time monitoring and intelligent management of ship circuit faults, and significantly improves the accuracy and response speed of fault detection. In the future, with the continuous progress of sensor technology and data acquisition systems, ship circuit fault detection and diagnosis will develop in the direction of higher intelligence and automation. Combining deep learning algorithms with time series data analysis may further improve the accuracy and efficiency of fault identification.

References

1. Huo, Y.: Application of RBF neural network in fault diagnosis of ship analog circuit. *Naval Science and Technology* **46**(10), 182–185 (2024)
2. Wang, Z., Fu, W., Liu, X.: Arc fault detection and hazard assessment of ship power system. *Shipbuilding and Offshore Engineering* **40**(3), 29–37 (2024)
3. Liu, R., Xu, C.: Accurate acquisition system for low-power electronic circuit faults of ship mainframe. *Ship Science and Technology* **41**(12), 97–99 (2019)
4. Liu, Y., Zhou, X., Wu, X.: Ship electronic fault classification and cataloging system based on RFID technology. *Naval Science and Technology* **45**(13), 174–177 (2023)
5. Yang, S., Xiang, X., Li, Z., Wang, D., Hou, Y.: Virtual simulation experiment for fault diagnosis of marine motor start-up control circuit. *Experimental Science and Technology* **19**(5), 48–53 (2021)
6. Tsaganos, G., Nikitakos, N., Dalaklis, D., et al.: Machine learning algorithms in shipping: improving engine fault detection and diagnosis via ensemble methods. *WMU J. Marit. Aff.* **19**(1), 51–72 (2020)
7. Ellefsen, A.L., Han, P., Cheng, X., et al.: Online fault detection in autonomous ferries: using fault-type independent spectral anomaly detection. *IEEE Trans. Instrum. Meas.* **69**(10), 8216–8225 (2020)

8. Cheliotis, M., Lazakis, I., Cheliotis, A.: Bayesian and machine learning-based fault detection and diagnostics for marine applications. *Ships and Offshore Structures* **17**(12), 2686–2698 (2022)
9. Orhan, M., Celik, M.: A literature review and future research agenda on fault detection and diagnosis studies in marine machinery systems. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment* **238**(1), 3–21 (2024)
10. Jiang, Y., Lan, G., Zhang, Z.: Ship engine detection based on wavelet neural network and FPGA image scanning. *Alex. Eng. J.* **60**(5), 4287–4297 (2021)
11. Han, P., Ellefsen, A.L., Li, G., et al.: Fault detection with LSTM-based variational autoencoder for maritime components. *IEEE Sens. J.* **21**(19), 21903–21912 (2021)
12. Maqsood, A., Oslebo, D., Corzine, K., et al.: STFT cluster analysis for DC pulsed load monitoring and fault detection on naval shipboard power systems. *IEEE Trans. Transport. Electrification* **6**(2), 821–831 (2020)
13. Wang, S., Dehghanian, P.: On the use of artificial intelligence for high impedance fault detection and electrical safety. *IEEE Trans. Ind. Appl.* **56**(6), 7208–7216 (2020)
14. Abid, A., Khan, M.T., Iqbal, J.: A review on fault detection and diagnosis techniques: basics and beyond. *Artif. Intell. Rev.* **54**(5), 3639–3664 (2021)
15. Ju, Y., Tian, X., Liu, H., et al.: Fault detection of networked dynamical systems: a survey of trends and techniques. *Int. J. Syst. Sci.* **52**(16), 3390–3409 (2021)



The Process of Building Color Extraction is Optimized with K-means Clustering Algorithm

Jian Liu and Junru Chen^(✉)

Shenyang Jianzhu University, Shenyang 110168, China
1223718524@qq.com

Abstract. In the field of architectural decoration, the importance of color matching is increasingly prominent. It not only affects the aesthetics of space, but also relates to people's emotions and behavior. This study aims to explore the role of CNN in color matching in architectural decoration and evaluate its effectiveness. This study uses Convolutional Neural Network (CNN) as the core algorithm and combines it with K-means clustering algorithm to optimize the color extraction process. By collecting and preprocessing color data of building decoration materials, this study trains a CNN model and validates its performance using historical color matching case data. The experimental results show that the average value of the CNN model on the color beauty index is 0.955, which is better than the support vector machine (SVM)'s 0.814, demonstrating the generalization ability and accuracy of CNN in color matching tasks. CNN also performs well in originality ratings, with an average score of 9.465, higher than SVM's 7.74, demonstrating its ability to generate high-altitude creative color schemes. However, there are limitations to the research, including limitations in dataset size and diversity, as well as the need to improve algorithm speed and ability to handle large-scale datasets. Future research will focus on expanding datasets, optimizing algorithm performance, and exploring the potential applications of intelligent algorithms in other design fields.

Keywords: Color Matching · Convolutional Neural Network · K-means Clustering · Color Beauty

1 Introduction

In the field of architectural decoration, the importance of color matching is self-evident. It not only affects the aesthetics and practicality of the space, but also directly relates to the emotions and behaviors of the residents. With the rapid development of artificial intelligence technology, the application of intelligent algorithms in the field of design is gradually increasing, especially in color matching, which has shown great potential. Intelligent algorithms can learn the rules of color matching by analyzing historical data, providing designers with scientific and objective decision support. However, how to effectively apply intelligent algorithms to color matching in architectural decoration and improve the diversity and innovation of color matching is still a problem worthy of in-depth research.

This article aims to explore the role of intelligent algorithms in color matching in architectural decoration, evaluate their effectiveness in practical applications, and propose improvement directions. This study collects and processes color data of building decoration materials, and trains a model capable of automatically generating color matching schemes using convolutional neural networks (CNN) and K-means clustering algorithm. By comparing with support vector machine (SVM), this study evaluates the performance of CNN in terms of color beauty and originality.

The article first reviews relevant work and explores the current application status of intelligent algorithms in color matching; next, the article introduces the methods of this study, including data collection, design and implementation of intelligent algorithms, model training and validation; then, the article demonstrates the process of generating and optimizing color matching schemes; finally, in the results and discussion section, the article analyzes the experimental results, discusses the potential application of intelligent algorithms in building decoration color matching, and proposes future research directions.

2 Related Work

In the study of color matching and its application in different fields, scholars have conducted in-depth analysis and discussion from historical works of art to modern design practice, from natural environment to man-made environment. Li Hexiao [1] studied the scroll paintings in the Tang Dynasty and found that the color collocation of figure costumes pays attention to the use of contrast colors and embellishment colors, and is good at matching bright colors such as red, yellow, cyan and green with black and white, which makes the color collocation of figure costumes more abundant. Xu [2] used the methods of literature review, case analysis, summary and induction to analyze the development of garden color, the change and configuration of plant color and the application of color psychology theory, and put forward some conclusions, such as the design conforms to the theme, the rational application of four seasons colors, and the principle of integrity, in order to provide some reference value for garden workers in the future plant color configuration, so that plant color can be better applied in landscape design. Li [3] found that the color design of furniture involves many aspects, such as technology, material selection, furniture functionality, color practicality, environmental factors and ergonomics, etc. These elements are mutually restricted, and reflect the perfect combination of technology and art, as well as the integration of technology and modern aesthetic concepts. Chen Hesun [4] studied the color matching of indoor soft clothes, analyzed the psychological needs of people in different seasons from the natural colors of the four seasons, demonstrated the color matching scheme of indoor soft clothes under seasonal changes, and introduced the characteristics of seasonal changes into the color matching of indoor soft clothes, so as to make the color of soft clothes more harmonious and unified with the characteristics of seasonal changes and people's psychological needs under the condition that the overall indoor style is unchanged. Sun [5] explored the relationship between color and clothing design, analyzed the influencing factors and principles of color collocation in clothing design, and analyzed the color collocation in clothing design of different groups such as children, teenagers and youth according to their ages for reference.

Talib et al. [6] analyzed the implementation technology of artificial intelligence algorithm on hardware, including algorithm selection, hardware platform design and performance evaluation. Fatemidokht et al. [7] put forward a routing protocol based on artificial intelligence algorithm, which uses unmanned aerial vehicles to improve the routing efficiency and security in vehicle networks, and its performance is verified by simulation experiments. Balasubramaniam [8] adopted skin image data set, and constructed an artificial intelligence algorithm model through preprocessing, feature extraction and SVM classifier training. Seyyed-Kalantari et al. [9] analyzed the application of artificial intelligence algorithm in chest radiography diagnosis, and evaluated its diagnostic bias in different patient groups. Wehbe [10] developed an artificial intelligence algorithm called DeepCOVID-XR, which was used to detect COVID-19 in chest radiation images. Although the above research provides profound insights and practical application strategies in their respective fields, the existing research often focuses on the deepening of a single field, lacking the discussion on the comprehensive role of color matching in actual architectural decoration. By using intelligent algorithm, this article will explore how to combine the scientific and artistic color matching to achieve a more harmonious, beautiful and user-friendly indoor environment.

3 Method

3.1 Data Collection

When studying the function of intelligent algorithm in architectural decoration color matching, the color data of architectural decoration materials are collected first. This step involves collecting sample data of materials and color matching used in architectural decoration from various channels. These data may include interior decoration cases with different styles and different functional spaces and related color value information, as shown in Table 1:

Table 1. Color data

Material Type	Color Name	RGB Values	Reflectance	Application Area
Paint	White	255, 255, 255	0.91	Interior walls
Marble	Gray	128, 128, 128	0.32	Flooring, walls
Wood	Light Brown	190, 152, 122	0.5	Flooring, furniture
Glass	Colorless	0, 0, 0	0.78–0.82	Windows, decoration
Metal	Silver	192, 192, 192	0.88–0.99	Modern decoration

Then the OpenCV library is used for image semantic segmentation and color clustering analysis. OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library, which provides a variety of image processing and analysis tools. Through OpenCV, this study can carry out semantic segmentation on the collected architectural decoration images, and identify and distinguish

different objects and areas such as walls, furniture and decorations in the images. This study also applies color clustering analysis to identify and classify the main colors in the image, which helps to understand the application and distribution of colors in different building decoration materials.

3.2 Design and Implementation of Intelligent Algorithm

In this article, the convolutional neural network (CNN) is selected as the core algorithm [11, 12]. In the study of color matching in architectural decoration, CNN has learned complex patterns and relationships from a large number of color matching samples. CNN can automatically extract color features from images through its multi-layer convolution layer, pool layer and full connection layer, and these features can then be used in the decision-making process of color matching.

Let the color space of the input image be X , which is expressed as a vector, and each dimension of the vector represents the intensity value of a color channel. The convolution neural network extracts color features through the following steps:

Convolutional layer: X extracts features through convolutional layer C to obtain feature map F :

$$F = C(X) \quad (1)$$

Pool layer: the feature map F is reduced in dimension through pool layer P , which reduces the number of parameters and extracts important features:

$$G = P(F) \quad (2)$$

Fully connected layer: the reduced feature G is classified or regressed through the fully connected layer D to get the final color matching decision Y :

$$Y = D(G) \quad (3)$$

In order to optimize the color extraction process, this study combines K-means clustering algorithm. K-means is a clustering algorithm that can divide data points into a predetermined number of clusters. In color extraction, K-means algorithm can help this study to quantify the color space in the image and identify the main colors in the image. By setting the appropriate number of clusters K-means, the complex color space can be simplified into a set of representative colors, which can be used as the basis of the subsequent color matching scheme. The optimization of K-means algorithm not only improves the efficiency of color extraction, but also enhances the innovation and diversity of color matching schemes.

In the process of implementation, this study firstly preprocesses the collected architectural decoration image data by normalization and dimension reduction to meet the input requirements of CNN model. Then, this study inputs the preprocessed data into CNN model for training, and the model automatically adjusts its internal weight by learning the color features in the data to capture the most representative color information. In this study, the trained CNN model is used to extract the color features of the new architectural decoration image and get a set of color feature vectors.

Finally, this study takes these feature vectors as input, and uses K-means clustering algorithm to quantify the color space, so as to get a set of optimized color matching schemes. These schemes not only consider the aesthetic principles of color, but also combine the environmental factors and user preferences in practical application, providing users with scientific, objective and innovative suggestions on color matching.

3.3 Model Training and Verification

In the part of model training and verification, this study uses historical color matching case data to train CNN model [13, 14]. This step involves dividing the collected building decoration color data set into training set and test set. The training set contains a variety of architectural decoration color matching samples, which are used to guide CNN model to learn the law of color matching. In the training process, CNN model constantly adjusts its internal parameters through forward propagation and backward propagation algorithms to minimize the difference between the predicted color matching and the actual color matching. In this way, the model can gradually learn how to predict a harmonious and visually attractive color matching scheme according to the input color characteristics.

In this study, the performance of the model is verified by test set data. The test set contains color matching samples that the model has never seen before, and these samples are used to evaluate the generalization ability of the model. In the process of verification, this study calculated the accuracy and recall of the model on the test set to ensure that the model not only performs well on the training set, but also can be accurately extended to new data. This study also qualitatively analyzes the output of the model, displays the color matching schemes generated by the model in a visual way and invites professional interior designers to evaluate these schemes to ensure the practicality and innovation of the model.

3.4 Generation and Optimization of Color Matching Scheme

The generation and optimization of color matching schemes are achieved by using trained CNN models to generate color matching schemes [15]. In this step, the model is exposed to a new architectural decoration design case, and the model puts forward a preliminary color matching scheme for a given design situation according to the color matching rules it learned in the training stage. These schemes integrate the hue, lightness and saturation of colors to create a harmonious and visually impactful indoor environment.

This study iterates and optimizes the scheme according to the feedback from designers. In practical application, the professional opinions and aesthetic judgments of designers are very important. Therefore, a feedback mechanism is established in this study, which allows designers to evaluate and modify the color matching scheme generated by the model. The designer's feedback is used to fine-tune the parameters of the model or adjust the color matching rules, so as to make the scheme more in line with the actual design requirements and aesthetic standards. This iterative process involves several rounds of evaluation and optimization until the generated color matching scheme meets the designer's requirements.

K-means clustering algorithm is also used to optimize the color extraction process. Through this method, we can identify the key color themes from a large number of color

data, and then generate more creative and practical color matching schemes based on these themes. The application of K-means algorithm not only improves the efficiency of color extraction, but also enhances the diversity and innovation of color matching schemes.

4 Results and Discussion

4.1 Color Matching Scheme Color Beauty

When discussing the intelligent solution of architectural decoration color matching, CNN's performance in image recognition and processing has proved its potential in color matching. It can analyze and learn complex color patterns and provide innovative color schemes for designers. In order to comprehensively evaluate the application effect of CNN in this field, this study compares it with support vector machine (SVM) and compares the diversity of color matching schemes generated by them. This comparison will help this study to understand the unique advantages and potential limitations of CNN in dealing with the task of building decoration color matching. The comparative results of visual beauty index are shown in Fig. 1:

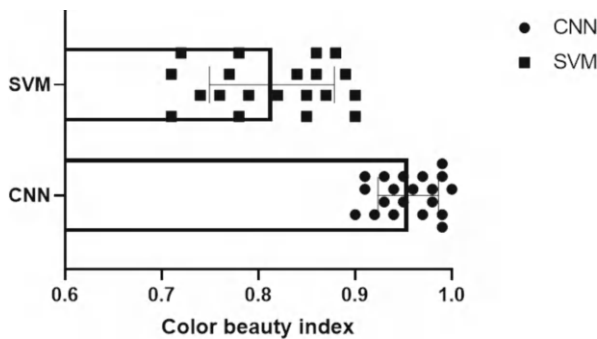


Fig. 1. Color Beauty Index

As shown in Fig. 1, based on the above data, this study makes a detailed data analysis on the performance of CNN and SVM in color beauty index. The analysis results show that the average color beauty index of CNN is 0.955, while the average color beauty index of SVM is 0.814, which indicates that CNN may have better generalization ability and accuracy in color matching tasks. In the comparison between the best performance and the worst performance, CNN's best color beauty index reaches 1 (appearing many times), while the worst color beauty index is 0.9, both of which are better than SVM's best color beauty index 0.9 and worst color beauty index 0.71 (appearing twice). Therefore, based on these data, this study concludes that CNN is a better choice for color matching tasks, which is superior to SVM in average performance, stability and best/worst performance.

4.2 Innovation of Color Matching Scheme

Innovation is an indispensable consideration when discussing the color matching scheme in the field of architectural decoration. It is not only related to the attraction and sense of the times of design works, but also directly affects the visual influence and emotional expression of space. Innovative color matching can break the traditional shackles and introduce novel design concepts, thus stimulating a unique aesthetic experience. The innovation of the scheme is measured by the originality scores of 10 different experts, and the results are shown in Fig. 2:

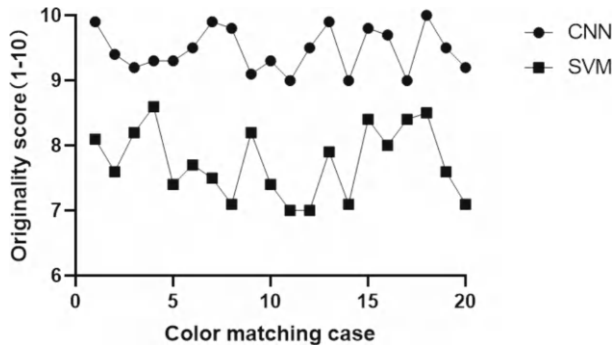


Fig. 2. Originality score

As shown in the data in Fig. 2, from the average point of view, the average value of color matching originality of CNN is 9.465, while the average value of SVM is 7.74. This significant difference shows that CNN tends to produce more original color schemes in color matching tasks, while SVM’s color schemes are less original. In all 20 samples, CNN scored higher than SVM in color matching originality. Especially in samples 1, 7, 13 and 18, the score of CNN is close to or reaches 10, while the score of SVM is relatively low, which further highlights the ability of CNN in generating highly original color schemes. To sum up, based on the data analysis provided, it can be concluded that CNN is more original and stable than SVM in color matching tasks.

4.3 Discussion

Through research, it is found that the harmony and contrast between different colors can be identified by using CNN training model, so as to generate a color matching scheme with visual impact. This discovery supports the original hypothesis that intelligent algorithms can improve the scientificity and objectivity of color matching. However, the research also found that the prediction accuracy of intelligent algorithm still needs to be improved when dealing with some complex or unconventional color matching. This is related to the diversity and representativeness of the data set, indicating that the preliminary hypothesis of this study needs further verification and adjustment.

Comparing these results with those found in other studies, we can find that although the application of intelligent algorithm in color matching is still in the initial stage, it

has shown great potential in improving design efficiency and innovation. Some studies have successfully realized the automatic analysis and evaluation of architectural colors through intelligent algorithms, which is consistent with the research goal of this study.

5 Conclusion

This article solves the problem of how to improve the scientificity, objectivity and innovation of color matching by using artificial intelligence technology by studying the application of intelligent algorithm in architectural decoration color matching. In this study, the convolutional neural network (CNN) is used as the core algorithm and K-means clustering algorithm is combined to optimize the color extraction process. Through experiments, this study verifies the superior performance of intelligent algorithm in color beauty and originality scoring. Compared with SVM, CNN shows higher accuracy and innovation ability.

Nevertheless, the scale and diversity of data sets in this study limit the test of model generalization ability; the running speed of the algorithm and the ability to deal with large-scale data sets need to be improved. These limitations suggest that it is necessary to further expand the data set and optimize the performance of the algorithm in the follow-up research.

Looking forward to the future, the application potential of intelligent algorithm in architectural decoration color matching is huge. Future research can explore the following new directions: expanding the size and diversity of the dataset to improve the generalization ability of the model; introducing parallel computing technology and optimizing algorithm implementation to improve the running speed and processing capability of algorithms; conducting comparative experiments by combining multiple machine learning algorithms or deep learning algorithms to explore algorithm models that are more suitable for fault diagnosis and prediction; further researching and developing automated fault response mechanisms, including intelligent fault isolation, evaluation, prioritization, and automated repair.

References

1. Ying, L., Xiao, H.: Study on color matching of figure costumes in scroll paintings in Tang Dynasty. *Dye. Finish. Technol.* **46**(4), 123–125 (2024)
2. Jinlan, X.: Discussion on the application of color matching in garden plants. *Modern Horticulture* **47**(1), 135–138 (2024)
3. Xingyi, L.: Color matching in furniture art design. *Footw. Technol. Desi.* **4**(3), 120–122 (2024)
4. Mo, C., Dazhi, S.: Study on the color matching of indoor soft clothes based on seasonal changes. *J. Dali Univ.* **8**(1), 80–84+F0003 (2023)
5. Haixia, S.: Discussion on color matching and application methods in modern clothing design. *Footw. Technol. Desi.* **3**(18), 6–8 (2023)
6. Talib, M.A., Majzoub, S., Nasir, Q., et al.: A systematic literature review on hardware implementation of artificial intelligence algorithms. *J. Supercomput.* **77**(2), 1897–1938 (2021)
7. Fatemidokht, H., Rafsanjani, M.K., Gupta, B.B., et al.: Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **22**(7), 4757–4769 (2021)

8. Balasubramaniam, V.: Artificial intelligence algorithm with SVM classification using der-mas-copic images for melanoma diagnosis. *J. Artif. Intel. Capsule Netw.* **3**(1), 34–42 (2021)
9. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., et al.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**(12), 2176–2182 (2021)
10. Wehbe, R.M., Sheng, J., Dutta, S., et al.: DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set. *Radiology* **299**(1), E167–E176 (2021)
11. Jacob, I.J., Darney, P.E.: Artificial bee colony optimization algorithm for enhancing routing in wireless networks. *J. Artif. Intel. Capsule Netw.* **3**(1), 62–71 (2021)
12. Waheed, S.R., Rahim, M.S.M., Suaib, N.M., et al.: CNN deep learning-based image to vector depiction. *Multimedia Tools and Appl.* **82**(13), 20283–20302 (2023)
13. Elngar, A.A., Arafa, M., Fathy, A., et al.: Image classification based on CNN: a survey. *J. Cybersecurity and Info. Manage.* **6**(1), 18–50 (2021)
14. AbdulAzeem, Y., Bahgat, W.M., Badawy, M.: A CNN based framework for classification of Alzheimer’s disease. *Neural Comput. Appl.* **33**(16), 10415–10428 (2021)
15. Kim, B., Yuvaraj, N., Sri Preethaa, K.R., et al.: Surface crack detection using deep learning with shallow CNN architecture for enhanced computation. *Neural Comput. Appl.* **33**(15), 9289–9305 (2021)



Research on Risk Monitoring and Early Warning Technology for Special Disaster Emergency Rescue Site Based on Reformer Model

Lei Zhang, Yufeng Fan^(✉), and Zhenpeng An

Shenyang Fire Research Institute of MPS, Shenyang, China
syjsml@163.com

Abstract. Special sites such as petrochemical plants and hazardous chemical production and storage often have toxic and harmful gas diffusion and spread in case of safety leakage accidents. How to ensure the safety of emergency rescue personnel is currently a research hotspot in the field of emergency rescue. This article proposes a toxic and harmful gas risk monitoring and early warning system suitable for fire emergency rescue sites. The system is based on the existing emergency communication system and equipment of China's fire rescue teams, and realizes real-time perception, transmission, and processing of toxic and harmful gas concentrations around the body area of individual firefighters. Through grid based rescue sites and matching calibration with positioning data, the Reformer model is applied to predict and warn the spread concentration of toxic and harmful gases in the grid area, can provide data support for scientific command by rescue site commanders.

Keywords: Special Site · Risk Monitoring And Warning · Reformer Model · Deep Learning

1 Introduction

In the daily production process of petrochemical plants, hazardous chemical production and storage sites, various chemical raw materials are involved, and flammable, explosive, and toxic gases are easily generated during the production process. Once a leakage accident occurs, it is highly likely to lead to major production safety accidents. Therefore, China has extremely high requirements for such enterprises or places, especially for the high-level regulatory requirements for the production, storage, transportation, and other links of such enterprises. According to Paper 2 of the Implementation Measures for the Safety Production License of Hazardous Chemical Production Enterprises in China (Order No. 41 of the State Administration of Work Safety), hazardous chemical production enterprises refer to enterprises that are legally established and have obtained a business license or industrial and commercial approval document to engage in the production of final or intermediate products listed in the Catalogue of Hazardous

Chemicals. Paper 3 of the Measures stipulates that enterprises shall obtain the Safety Production License for Hazardous Chemicals (hereinafter referred to as the Safety Production License) in accordance with the provisions of these Measures. Enterprises that have not obtained a safety production license are not allowed to engage in the production of hazardous chemicals. If an enterprise is involved in the use of toxic substances, in addition to the safety production license, it shall also obtain the occupational health and safety license in accordance with the law [1].

Although the country has put forward high requirements for the production safety of petrochemical plants and hazardous chemical production and storage enterprises, due to the complexity of the entire process and the multitude of uncontrollable factors, multiple major fire accidents have occurred in recent years, resulting in irreparable losses. At present, as the main force in the disposal of petrochemical and hazardous chemical fire accidents, the scientific disposal of such disasters is the key to avoiding casualties and reducing property losses. The premise of scientific disposal is to obtain various effective information on the scene in real time, such as the types of hazardous chemicals, ignition points/leakage sources, and the development trend of accidents. These information will directly affect the firefighting and rescue methods adopted by the firefighting and rescue teams, how to deploy the next step of rescue, and so on. Therefore, how to obtain various effective information on the scene through technical means and effectively use various information for risk assessment in the event of a disaster is an important problem that urgently needs to be solved.

2 Related Work

Traditional disaster rescue scene simulation mainly relies on the Fire Dynamics Simulator (FDS) software, which serves as a fundamental tool for fire dynamics and combustion studies. It is mainly used in smoke control design, detector start-up time simulation research, and fire reconstruction. The basic principles of the software are mainly based on fluid dynamics models, turbulence models, combustion models, radiation models, etc. FDS is currently mainly used for reviewing fire scenes, reconstructing the development process of fires, and analyzing the causes of fires. In the process of fire simulation, a large number of data sources need to be used as inputs, such as scene conditions, burning substances, quantity of substances, weather conditions at the time of fire occurrence, and so on.

Jiabin Gao et al. [2] studied the influence of different conditions on the development characteristics of container fires using a combination of full-scale experiments and FDS simulations, focusing on dangerous goods containers. This research work can provide support and assistance for the fire safety design, fire prevention, and control of closed containers. Su Feng et al. [3] constructed a dynamic risk assessment model for cable fires in utility tunnels based on different response situations of safety barriers such as sensors, sprinkler systems, ventilation systems, and fire doors. They achieved FDS simulation modeling to explore dynamic risk assessment and emergency decision-making optimization schemes for cable fires in utility tunnels under fire scenarios. Wang Shuo et al. [4] used the observation deck of Shanghai Jinmao Tower as the research object to explore the impact of fixed window to ground ratio simulation modeling using FDS on smoke

spread in the observation deck when mechanical smoke exhaust failure occurs. Based on the research conclusions, it is recommended to add a clause when revising the “Technical Standard for Building Smoke Control and Exhaust Systems” (GB51251–2017): the total area of fixed windows located near the exterior wall and not on the top floor area should not be less than 5% of the floor area. Liu Hongman [5] analyzed the causes of the “4.17” major fire accident in Wuyi, Jinhua. Through FDS reconstruction simulation of the fire scene, the occurrence process of the fire was reconstructed, and the parameters of the fire temperature, smoke, oxygen concentration, heat release rate, and various sections were analyzed. Combined with Pathfinder evacuation simulation, a comparative analysis was conducted on the relevant disaster causing factors of the fire accident. Wenqian L et al. [6]. Proposed a method for simulating hydrogen jet fires using fire dynamics simulation software FDS. To avoid modeling the actual nozzle, high-speed Lagrangian papers released from the virtual nozzle are introduced to simulate the released hydrogen gas. By comparing the simulation results with five existing experiments in a rectangular steel cabin with an opening, the ability of the FDS model to predict gas temperature was verified. He Hui et al. [7], Yan Weidong et al. [8], Chen Chen et al. [9] applied FDS to model and simulate the performance-based fire prevention, fire evacuation, and safety performance of exhibition centers, libraries, sports halls, and other venues. Lee Jaiho et al. [10] conducted a validation study on the Fire Dynamics Simulator (FDS) model for multi liquid pool fire scenarios that occur in multiple train compartments. Overall, due to the difficulty of quickly obtaining a large amount of information at disaster sites, it is difficult to use FDS for situational inference at disaster accident sites.

3 Design of Toxic and Harmful Gas Risk Monitoring and Early Warning System for Fire Emergency Rescue Sites

The operating environment of the toxic and harmful gas risk monitoring and early warning system for fire emergency rescue sites includes the fire individual domain network, fire combat network, and fire command and dispatch network. The system has three functional units: perception, transmission, and risk prediction of toxic and harmful gases in fire emergency rescue sites. The system schematic is shown in Fig. 1.

The toxic and harmful gas sensing unit operates within the fire soldier’s body area network, consisting of communication terminals held/worn by fire soldiers and gas sensors attached to individual protective clothing. The types of gas sensors include eight typical harmful gas as HCN , CO , HCl , HBr , HF , SO_2 , NO_2 , CH_2O s in a fire scene, as well as two combustible gases as CH_4 , H_2 , are collected as gas concentrations at that point in ppm. The sensor collects gas type and concentration information and wirelessly transmits it to the individual communication terminal. The individual communication terminal supports BeiDou satellite positioning and loads location information before wirelessly transmitting it to the firefighting network. The individual soldier body area network is a local area network within 10 m of the individual soldier, and data transmission between gas sensors and communication terminals can be carried out using Bluetooth and Wi Fi radio frequency. The firefighting network is a local area network that covers the entire combat site. Currently, fire teams mostly use Mesh strategy for wireless networking, with communication frequencies of 600 MHz and 1.4 GHz. Communication terminals can

be directly or back-to-back connected to the firefighting network (i.e., a single soldier carries another firefighting network networking device, and the communication terminal accesses the networking device through Wi Fi).

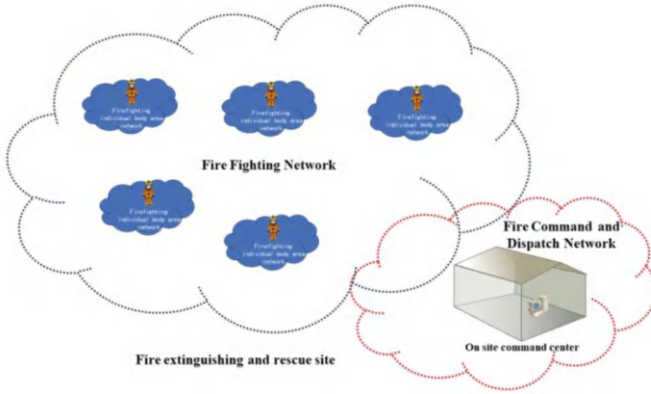


Fig. 1. System schematic diagram

The toxic and harmful gas transmission unit operates in the firefighting network, mainly composed of multiple firefighting individual communication terminals and multiple firefighting network networking devices. The initiating end is the firefighting individual communication terminal, and the receiving end is the firefighting individual's nearby firefighting network networking device. The firefighting network networking is a local area network based on Mesh strategy, and the firefighting network networking device can wirelessly access the fire command and dispatch network, with Mesh direct access as the access method.

The toxic and harmful gas risk prediction unit operates on hardware computing and storage devices within the on-site command center, and the operating network is the fire command and dispatch network. The fire command and dispatch network can be connected to the brigade level fire communication command center through wireless public/private networks to achieve data sharing and interconnection.

4 System Input Design

4.1 Scene Segmentation

Define the fire emergency rescue site as $Feilds$, divide the site into grids of $1m \times 1m$, and use the GIS system of the fire brigade to select the center of the site as the origin. $Feilds_{(i,j)}$ represents the area of the site with coordinates (i,j) and an area of $1m \times 1m$.

4.2 Data Acquisition

The concentration attribute of toxic and harmful gases in the $Feilds_{(i,j)}$ area of the fire emergency rescue site is the $Vector_{gas}$ vector,

$$Vector_{gas} = [HCN, CO, HCl, HBr, HF, SO_2, NO_2, CH_2O, CH_4, H_2] \quad (1)$$

The internal elements are the concentrations of 8 typical harmful gases as HCN , CO , HCl , HBr , HF , SO_2 , NO_2 , CH_2O , and 2 combustible gases as CH_4 , H_2 in ppm. If the value is 0, it means that the gas does not exist at that location.

4.3 Data Input Model

Define the data matrix for monitoring toxic and harmful gas risks at fire emergency rescue sites as $GasMatrix_{input}$,

$$GasMatrix_{input} = [Vector_{gas(i,j)}] \quad (2)$$

The matrix elements represent the concentrations of 8 harmful gases and 2 flammable gases in ppm. If the value is 0, it means that the gas does not exist at that location.

The concentration of toxic and harmful gases at the fire emergency rescue site changes over time, which is a complex process influenced by multiple factors. It is related to the location of the leakage source, the inventory of the leakage source, wind direction, wind speed, temperature, humidity, air pressure, etc. However, these related factors are difficult to obtain in real time during disasters and accidents. Moreover, FDS software based on fluid mechanics is suitable for laboratory modeling, and the accuracy of simulation and deduction for the site needs to be discussed. Based on these analyses, in this paper, we did not incorporate environmental factors such as wind direction, wind speed, temperature, humidity, and air pressure into the input of the prediction model for gas diffusion. Instead, we extracted trend features through machine learning models based on the dynamic changes in time-series data, which were then used to predict the concentration changes of toxic and harmful gases in the region.

5 Model Building

The Reformer model is an efficient attention mechanism model suitable for processing long sequence data. This model is used to predict the concentration changes of toxic and harmful gases in various areas of the fire emergency rescue site. Its workflow includes steps such as data preprocessing, model training, and predictive output.

5.1 Data Preprocessing

Before building a predictive model, it is necessary to preprocess the collected toxic and harmful gas data. The preprocessing steps are as follows:

- Data normalization: Normalize the concentration data of toxic and harmful gases in each region to ensure that the data is on the same scale, facilitating model training.
- Time series partitioning: Normalize the data into training and testing sets according to the time series, ensuring that the model can learn the temporal features of the data.

The input matrix after data normalization is denoted as,

$$NormGasMatrix_{input} = [NormVector_{gas(i,j)}] \quad (3)$$

Among them, $NormVector_{gas(i,j)}$ includes concentration data of toxic and harmful gases.

5.2 Model Training

- **Model initialization:** During the model initialization phase, multiple hyperparameters need to be set to construct the Reformer model, including the number of layers, attention heads, hidden units, sequence length, and embedding dimension. The number of layers is usually between 6 and 12, with each layer containing a self attention module and a feedforward neural network module. The number of attention heads is usually 8 to 16, with each head independently capturing different features in the input sequence and merging them together. The number of hidden units is usually between 512 and 2048, used for processing complex feature representations. The length of the sequence determines the number of time steps that the model can handle, and Reformer uses local sensitive hashing (LSH) technology to process long sequence data to reduce computational complexity. The embedding dimension is usually the same as the number of hidden units, determining the size of each input feature vector. By setting these hyperparameters, the model can adapt to different task requirements and achieve efficient time series data processing.
- **Definition of loss function:** The loss function is defined as the mean square error (MSE) between the predicted value and the true value.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (4)$$

where N is the number of samples, \hat{y}_i is the predicted value of the i -th sample, and y_i is the true sample value.

- **Model training:** Model training includes three steps: model initialization, loss function definition, and model training. When initializing the Reformer model, hyperparameters such as the number of layers, attention heads, hidden units, sequence length, and embedding dimension need to be set. The loss function is defined as the mean square error (MSE) between the predicted value and the true value, which measures the accuracy of the model's predictions. During the training process, the normalized training set data is first input into the model in batches, and the predicted values are calculated through forward propagation. The loss function is then used to calculate the error. Then, the gradient is calculated using the backpropagation algorithm, and the model parameters are updated using optimization algorithms such as Adam. The entire training process involves multiple iterations, constantly adjusting model parameters to minimize the loss function value, and evaluating model performance using a validation set, adjusting hyperparameters if necessary. Ultimately, the Reformer model is able to learn the time series characteristics of toxic and harmful gas concentrations, improving the accuracy and reliability of predictions.

5.3 Model Prediction

After training, use the test set data for prediction and obtain the concentration values of toxic and harmful gases for future time steps. The prediction process is as follows:

- **Input data:** Input the normalized test set data into the model.

- Output predicted value: The model outputs the concentration values of toxic and harmful gases in the predicted time series.
- Anti normalization processing: Perform anti normalization processing on the predicted values to restore them to the original concentration scale.

The output of the Reformer model includes the predicted concentration values of toxic and harmful gases in each region based on time series. The prediction matrix output by the model is represented as follows:

$$PredictedGasMatrix_{output} = [PreVector_{gas(i,j,t)}] \quad (5)$$

Among them, $PreVector_{gas(i,j,t)}$ represents the predicted concentration of toxic and harmful gases in region (i,j) at time t .

6 System Warning Process

In the field of real-time monitoring and early warning of toxic and harmful gases in petrochemical plants, the national standard GB/T 50493–2019 “Design Standard for Detection and Alarm of Combustible and Toxic Gases in Petrochemical Industry” is mainly used to ensure the personal safety and production safety of petrochemical enterprises, monitor the leakage of combustible or toxic gases in the production process and storage and transportation facilities, and timely alarm to prevent personal injury and fire and explosion accidents. Due to the fact that the application scenario of this paper is the emergency rescue site of fire rescue teams, mainly involving petrochemical plants, hazardous chemical production and storage places, etc. By monitoring the types and concentrations of toxic and harmful gases on site in real time, toxic and harmful gas monitoring information guarantee is provided for the rescue process of individual fire-fighters. Therefore, the first and second level alarm functions and parameters in GB/T 50493–2019 are not applicable to the present invention patent.

The Emergency Response Planning Guidelines (ERPG) published by the American Industrial Hygiene Association (AIHA) provide a detailed regional hazard classification principle based on the types and concentrations of toxic and harmful gases on site:

- ERPG-1: The maximum concentration of chemicals in the air that a person can be exposed to for an hour without developing any symptoms;
- ERPG-2: The concentration of chemicals in the air that a person can be exposed to for one hour without causing irreversible or serious health effects that would render them unable to take protective measures;
- ERPG-3: The concentration of chemicals in the air that a person can be exposed to for an hour without causing life-threatening effects.

In addition to the above categories, IDLH (Immediately Dangerous to Life and Health) is also defined as the maximum allowable concentration at which individuals are exposed to toxic gases for 30 min, have the ability to escape, and do not experience adverse symptoms or irreversible health effects.

6.1 Warning Judgment Method

Therefore, in this paper, the emergency rescue site area of the fire brigade is divided into four levels. ERPG-1 corresponds to individual firefighters who can carry out emergency rescue in this area without wearing protective equipment; ERPG-2 requires individual firefighters to wear protective equipment for emergency rescue in the area; ERPG-3 requires individual firefighters to wear protective equipment, and the length of emergency rescue time in the area should not exceed 1 h; Even if the IDLH fire brigade is wearing protective equipment, they should immediately exit the area.

Establish a warning model for individual firefighters participating in emergency rescue operations on site, defined as,

$$firefighter_i = level_j \quad (6)$$

In the above formula, *firefighter* represents the *i*-th firefighting soldier participating in the battle on site, where *i* is less than the total number of soldiers on site *n*, *firefighter_i* represents the warning level of the *i*-th firefighting soldier's location, *level_j* represents the warning level of the *j*-th level, where *j* is less than 4, *level₀* corresponds to the concentration of 10 toxic and harmful gases carried by the soldier within the concentration range specified by ERPG-1, *level₁* corresponds to the concentration of 10 toxic and harmful gases carried by the soldier within the concentration range specified by ERPG-2, *level₂* corresponds to the concentration of 10 toxic and harmful gases carried by the soldier within the concentration range specified by ERPG-3, *level₃* corresponds to the concentration of 10 toxic and harmful gases carried by individual soldiers reaching the IDLH specified concentration or exceeding the ERPG-3 specified concentration.

Timing should be carried out under the current warning level of a single soldier. When it exceeds 1 h, the existing warning level should be raised by 1 level. For example, when a single soldier's area is at *level₀* and the timing reaches 1 h, the current warning level should be raised from *level₀* to *level₁*.

Special regulations stipulate that when the warning level of the individual soldier's area changes, the timing should restart.

6.2 Warning Process

The warning process is shown in Fig. 2.

7 Conclusion

This paper proposes a toxic and harmful gas risk monitoring and early warning system for fire emergency rescue sites. Based on the existing emergency communication system and equipment of fire rescue teams, the system realizes real-time perception, transmission, and processing of toxic and harmful gas concentrations around the body area of individual firefighters. Through grid based rescue sites and matching with positioning data, the Reformer model is applied to predict and warn the spread and concentration of toxic and harmful gases in the grid area, providing data support for scientific command by on-site commanders and effectively ensuring the safety of firefighters on site.

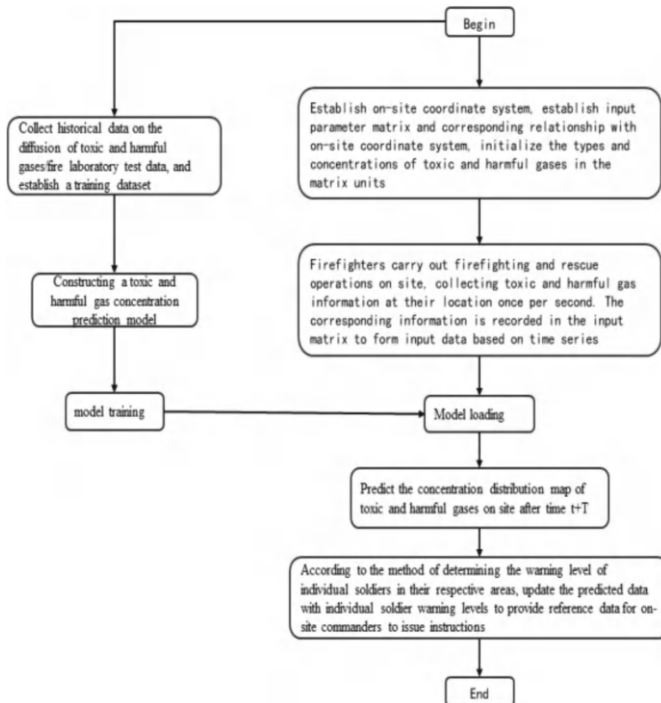


Fig. 2. System Warning Process Diagram

Different from the traditional toxic and harmful gas concentration spread model based on fluid dynamics FDS, this paper abandons the adverse effects of real-time acquisition of leakage source types, quantities, and on-site climate and environmental data on prediction and early warning at disaster sites. Instead, a method based on machine learning model is established to predict trends based on time-series data, which has higher practicality in real emergency rescue processes.

In this paper, the traditional method of information perception, transmission, and backend processing based on public wireless mobile communication networks is abandoned, and a method of integrating the fire individual domain network, fire combat network, and fire command and dispatch network is established. This method has higher applicability in the application of fire emergency rescue teams, especially in petrochemical plant areas, hazardous chemical production and storage sites, and other fire extinguishing and rescue sites with strong adaptability.

Acknowledgment. This work has been supported by The National Key R&D Program of China (2022YFC3006104).

References

1. National Administration of Work Safety, National Coal Mine Safety Supervision Bureau. Implementation Measures for Safety Production License of Hazardous Chemical Production Enterprises (2017)
2. Gao, J., Chen, B., Zhang, L., et al.: Development characteristics of container confined space fires based on FDS simulation and experimental. *J. Thermal Anal. Calorimet.* (2024). (prepubulish)
3. Feng, S., Jun, Z., Xiaoping, Z., et al.: Research on dynamic risk assessment of cable fire in pipe gallery based on FDS]. *China Safety Prod. Sci. Technol.* **20**(05), 77–83 (2024)
4. Shuo, W., Li, Y.: Study on the influence of fixed window to floor ratio fire smoke spread in sightseeing floors based on FDS. *Fire Protection Industry (Electronic Version)* **10**(03), 40–42 (2024). <https://doi.org/10.16859/j.cnki.cn12-9204/tu.2024.03.043>
5. Liu, H.: Numerical reconstruction analysis of the “4.17” major fire accident in Wuyi, Jinhua based on FDS. *Fire Sci. Technol.* **42**(12), 1733–1737 (2023)
6. Wenqian, L., Frank, M., Luisa, G., et al.: Analyzing the gas temperature of a hydrogen jet fire in a compartment with the Fire Dynamics Simulator. *Int. J. Hydrogen Ener.* 53 (2024)
7. Hui, H., Xin, H., Qiang, K., et al.: Research on performance based fire protection design of exhibition halls in large convention and exhibition centers based on FDS simulation. *J. Nankai Univ. (Natural Science Edition)* **56**(05), 85–89 (2023)
8. Weidong, Y., Zhikai, R., Xiaolei, W.: Study on library fire evacuation simulation based on FDS. *J. Shenyang Jianzhu Univ. (Social Sciences Edition)* **25**(05), 494–501 (2023)
9. Chen, C., Song, G.: Simulation and verification analysis of fire safety design for gymnasium based on FDS. *J. Zaozhuang Univ.* **40**(05), 113–119 (2023)
10. Jaiho, L., Byeongjun, K., Sangkyu, L., et al.: Validation of the fire dynamics simulator (FDS) model for fire scenarios with two liquid pool fires in multiple compartments. *Fire Safety Journal* 141 (2023)



Research on Intelligent Optimal Scheduling Algorithm for Vehicle Exhaust Emission in Railway Transportation System

Zixiang Xu¹(✉), Xiaokai Zhou², and Yishan Wang³

¹ School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
18360611797@163.com

² Development of Energy and Power Engineering, Tsinghua University, Beijing 100084, China

³ School of Mechanical and Vehicular Engineering, Beijing Institute of Technology,
Beijing 100081, China

Abstract. Railway transportation system faces the problem of tail gas emission while meeting the transportation demand, which puts forward higher requirements for energy saving and emission reduction. In order to meet this challenge, this paper proposes an intelligent optimal scheduling method based on improved genetic algorithm, and constructs a multi-objective optimization model to balance the transportation efficiency and tail gas emission. The effectiveness of the algorithm in reducing tailpipe emissions, improving transportation efficiency and shortening calculation time is verified through simulation experiments of actual railroad transportation network. The results show that the method has good optimization performance and practical application value, providing technical support for the green development of railroad transportation system.

Keywords: Railroad Transportation System · Intelligent Scheduling · Genetic Algorithm

1 Introduction

As an important part of modern transportation, railroad transportation system has been widely concerned about its high efficiency and environmental friendliness. With the continuous growth of transportation demand, the negative impact of tailpipe emissions from railroad vehicles on the environment is increasingly visible, and how to reduce emissions while improving transportation efficiency has become a key issue. Aiming at the limitations of the existing scheduling scheme in tailpipe emission optimization, this paper proposes an intelligent scheduling method based on improved genetic algorithm, and constructs a multi-objective optimization model by combining the actual operation data in order to balance the transportation efficiency and environmental protection needs. The study provides theoretical support and technical path for the green development of the railroad transportation industry, and expands new directions for the research and practice of related intelligent optimization scheduling problems.

2 Related Research

The study of intelligent optimal scheduling algorithm for tail gas emission in railroad transportation system involves several research fields, including tail gas emission characteristics, optimal scheduling model design, application of intelligent optimization algorithm and multi-task scheduling.

Ruchika Bhakhar and Rajender Singh Chhillar (2024) propose a dynamic multi-criteria scheduling algorithm for smart home tasks in fog-cloud IoT systems, focusing on optimizing performance and reducing resource consumption. This approach could be adapted to railway emission control, where dynamic scheduling could manage emissions based on varying train schedules and conditions [1]. Nikhil D. Khedkar et al. (2024) study engine control variables in a methane-diesel dual fuel engine, showing how optimization can reduce emissions. Similar strategies could optimize diesel-electric locomotive engines to reduce emissions and improve operational efficiency in railway systems [2]. Zexuan Han et al. (2024) investigate intelligent resource allocation for heavy vehicle safety. Such strategies could be adapted to railway systems to optimize energy consumption and emissions while ensuring locomotive stability [3]. Nian Liu and Yuehan Zhao (2024) present intelligent optimization techniques for reducing losses in electrical networks. In railway systems, similar strategies could optimize locomotive operations, improving energy efficiency and reducing emissions [4]. Ali Boroumand et al. (2024) introduce a heuristic scheduling algorithm for cloud computing to minimize costs. This concept can be applied to railway systems, optimizing locomotive scheduling to reduce idle times and fuel consumption, thereby minimizing emissions [5]. Arash Deldari and Alireza Holghinezhad (2024) explore IoT-based task scheduling for deadline-sensitive applications. Real-time scheduling in railway transport could help minimize emissions by optimizing train flow and reducing station waiting times [6]. Maciej Siedlecki et al. (2024) analyze how vehicle comfort systems influence emissions. Similar considerations could be applied to improve HVAC systems in trains, reducing energy use and emissions [7]. Caius Panoiu et al. (2024) examine real-time video processing for monitoring pantograph-catenary systems, a technology that could be used to track emissions and optimize train schedules for lower environmental impact [8]. Yongsheng Zhu et al. (2024) discuss privacy-preserving AI models in intelligent railway systems, which could help optimize emission control while securely processing environmental data [9]. Mohammad Ishaq et al. (2024) review optimization techniques for energy in suburban rail systems, suggesting that improving energy efficiency can reduce emissions, aligning with intelligent scheduling efforts [10]. Finally, Sajanraj Thandassery et al. (2024) focus on operational pattern forecasting in metro rail systems. This could be extended to emission forecasting, optimizing emissions during peak periods and high-traffic areas [11].

3 Intelligent Optimal Scheduling Algorithm

3.1 Problem Modeling

The core problem of optimal scheduling of vehicle exhaust emissions in the railroad transportation system is to construct a multi-objective optimization model, taking the minimization of exhaust emissions and the maximization of transportation efficiency as

the optimization objectives, while considering the actual constraints. Let there are N trains in the railroad transportation system, and the tailpipe emission of each train is Eq. (1).

$$E = f(S_i, v_i, F_i) \quad (1)$$

where S_i is the mileage of the train, v_i is the speed of the train, and F_i is the fuel consumption rate.

The objective function is Eq. (2):

$$E = \sum_{i=1}^N E_i = \sum_{i=1}^N f(S_i, v_i, F_i) \quad (2)$$

At the same time, the transportation efficiency of the train can be expressed as the ratio of the total volume of goods transported to the time (3):

$$\eta = \frac{\sum_{i=1}^N Q_i}{T} \quad (3)$$

where Q_i is the freight volume of train i and T is the total running time.

The constraints include the following: (1) time window constraints for train schedules to ensure that trains depart and arrive on time; (2) track capacity constraints to avoid multiple trains from occupying the same track at the same time; and (3) train energy consumption constraints to limit the traction output from exceeding the specified value.

In this paper, the modeling is validated using actual railroad system data, using 10 major lines and 100 train operation data of a railroad section. The model solving uses an improved intelligent optimization algorithm to ensure the feasibility and efficiency of the results.

3.2 Algorithm Design

The algorithm first generates an initial population through random initialization, and each individual represents a train scheduling scheme, including the train's departure time, speed and line selection. The population size is set to 100 to ensure the breadth of the search space. The fitness function is calculated by comprehensively evaluating tailpipe emissions and transportation efficiency, where tailpipe emissions are weighted higher to reinforce the energy saving objective. The selection operation uses a roulette method based on the fitness ratio to retain high-quality individuals to the next generation. The crossover operation generates a new scheduling scheme by single-point crossover, with the crossover probability set to 0.8 to balance the population diversity with the convergence speed of the algorithm. The mutation operation introduces small probability (0.05) scheme adjustments, such as changing train speed or departure time, for jumping out of the local optimal solution. In order to improve the convergence efficiency, the algorithm introduces a dynamic convergence determination strategy, which terminates the algorithm and outputs the current optimal solution when the standard deviation of the population's fitness is less than a set threshold for five consecutive generations. In the experiment, for the scheduling problem of 10 lines and 100 trains, the algorithm converges within 500 generations, which significantly reduces tailpipe emissions and improves transportation efficiency.

3.3 System Architecture

The system architecture designed in this paper consists of a data collection module, an algorithm optimization module, and a scheduling output module to achieve intelligent optimal scheduling of vehicle exhaust emissions in the railroad transportation system. In the first step, the data acquisition module collects the operation data of the railroad transportation system in real time, including the running speed of trains, fuel consumption, line occupancy status and environmental conditions. The module adopts a combination of embedded sensors and GPS system to ensure high accuracy and real-time data, and the sampling frequency is set to 1 time per second to meet the high-frequency scheduling requirements. In the second step, the algorithm optimization module receives the real-time data provided by the data acquisition module and removes the outliers through preprocessing, and then inputs the data into the improved genetic algorithm model. The algorithm uses a multi-objective optimization method to generate the optimal scheduling plan, including the train's running speed, departure time and route selection, and the optimization results are stored in the task queue. In order to accelerate the computation, the optimization module is deployed on a high-performance computing server, which supports parallel processing of scheduling optimization for multiple trains. In the third step, the scheduling output module transforms the optimized scheme into executable commands and transmits them to the train control system. The module maintains a real-time connection with the train through wireless communication technology to ensure the timely transmission and execution of instructions. The whole system architecture has been experimentally proven to have high reliability and real-time performance, which can effectively reduce exhaust emissions and improve transportation efficiency. The details are shown in Fig. 1.

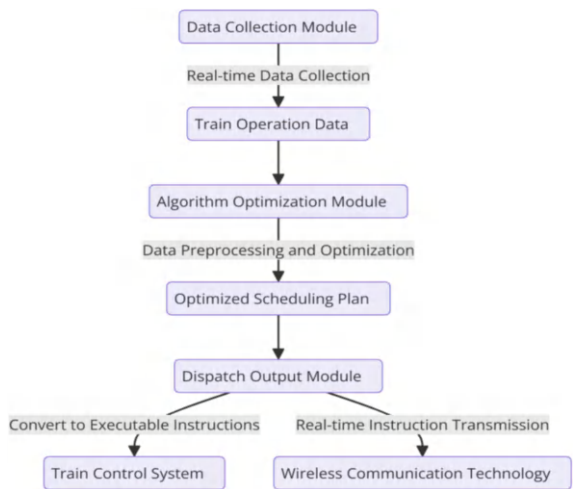


Fig. 1. System architecture flowchart

4 Results and Discussion

4.1 Experimental Setup

In order to verify the effectiveness of the proposed intelligent optimal scheduling algorithm, an actual railroad transportation network is selected for simulation. The network contains 10 major lines with a total length of about 2000 km, involving 100 freight trains. The initial parameters of the trains, such as load capacity, fuel consumption rate and running speed, are set based on real data. The experiments were conducted on a server with Intel Xeon 2.6 GHz processor and 32 GB RAM, and the algorithms were implemented using MATLAB R2021a. The evaluation metrics include tailpipe emissions, average transportation time and fuel consumption.

4.2 Simulation Experiment Results

By running the improved genetic algorithm, the optimized train scheduling scheme was obtained. The results show that compared with the traditional scheduling scheme, there is a significant improvement in tail gas emissions and transportation efficiency. The specific data are shown in Table 1.

Table 1. Comparison of the performance of different scheduling schemes

Scheduling plan	Exhaust emissions (t)	Average transportation time (h)	Fuel consumption (t)
Traditional scheduling scheme	500	50	1000
Improved genetic algorithm scheme	425	44	850

As can be seen from Table 1, after using the improved genetic algorithm, the tailpipe emission is reduced by 15%, the average transportation time is shortened by 6 h, and the fuel consumption is reduced by 15%. This shows that the optimized scheduling scheme has a significant effect in terms of energy saving and emission reduction and improving transportation efficiency. The details are shown in Fig. 2.

According to Fig. 2, it can be seen that the results of the comparison between the traditional scheduling scheme and the improved genetic algorithm scheme in terms of tailpipe emissions, average transportation time and fuel consumption are demonstrated. The figure clearly shows the advantages of the improved genetic algorithm in each index.

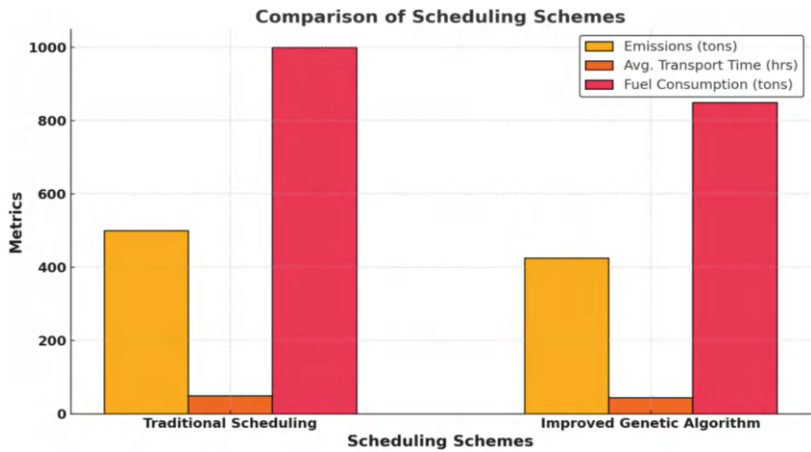


Fig. 2. Comparison of scheduling schemes

4.3 Comparison Analysis

In order to further verify the superiority of the algorithm, the improved genetic algorithm of this paper is compared with the traditional genetic algorithm and dynamic programming algorithm. The results are shown in Table 2.

Table 2. Performance comparison of different algorithms

Algorithm	Exhaust emissions (t)	Calculation time (s)	Transportation efficiency (t/h)
Dynamic programming algorithm	480	1200	2000
Traditional genetic algorithm	450	800	2100
Improved genetic algorithm	425	600	2200

Table 2 shows that the improved genetic algorithm outperforms the other two algorithms in terms of tailpipe emissions, computation time and transportation efficiency. Among them, the tail gas emission is reduced by 11.5%, the computation time is shortened by 50%, and the transportation efficiency is improved by 10% than the dynamic programming algorithm. This shows that the algorithm in this paper has a comprehensive advantage in optimization performance and computational efficiency. The details are shown in Fig. 3.

According to Fig. 3, it can be seen that the comparison results of dynamic programming algorithm, traditional genetic algorithm and improved genetic algorithm in terms of tailpipe emission, calculation time and transportation efficiency. The curves in the figure clearly reflect the performance advantages of the improved genetic algorithm.

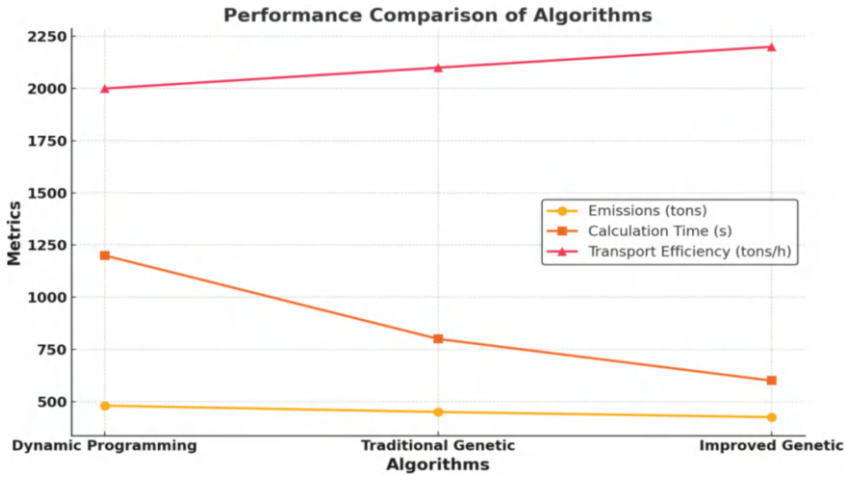


Fig. 3. Comparison of the performance of different algorithms

4.4 Discussion of Results

The experimental results show that the improved genetic algorithm has significant advantages in optimizing railroad transportation scheduling. Compared with the traditional method, the algorithm performs well in three aspects: tail gas emission, transportation efficiency and calculation time. The tail gas emission is reduced to 425 tons (about 12% reduction), the transportation efficiency is improved to 2200 tons/hour, and the calculation time is reduced to 600 s. Through the analysis of the optimization scheme, the improved genetic algorithm effectively reduces empty train operation and congestion by dynamically adjusting the train speed and departure time. The balanced weights in the optimization objective function further enhance the adaptability of the algorithm to multi-objective optimization. From the convergence curve of the experimental results, the algorithm reaches stability within 200 generations, proving its convergence speed and stability.

5 Conclusion

The research in this paper proposes a railroad transportation scheduling optimization method based on improved genetic algorithm, which achieves significant results in reducing tailpipe emissions, improving transportation efficiency and shortening calculation time. Through experimental verification, the optimization scheme effectively reduces fuel consumption and exhaust emissions, and significantly improves the scheduling efficiency and stability of the system. In the future, the real-time environmental data and dynamic adjustment mechanism should be further integrated to explore the multi-objective optimization method under more complex scenarios, focusing on the computational efficiency and adaptability of the algorithm, so as to provide more reliable technical support for the green development of large-scale railroad transportation system.

References

1. Bhakhar, R., Chhillar, R.S.: Dynamic multi-criteria scheduling algorithm for smart home tasks in fog-cloud IoT systems. *Sci. Rep.* **14**(1), 29957 (2024). <https://doi.org/10.1038/s41598-024-81055-0>
2. Khedkar, N.D., Sarangi, A.K., Sreedhara, S.: Impact of engine control variables on low load combustion efficiency and exhaust emissions of a methane-diesel dual fuel engine. *Proc. Inst. Mech. Eng., Part D: J. Automobile Eng.* **238**(14), 4551–4568 (2024). <https://doi.org/10.1177/09544070231197613>
3. Han, Z., Yongjie, L., Li, H., Zhang, J.: Research on intelligent optimal allocation control strategy for stability and braking of heavy tyre blowout vehicle. *Proc. Instit. Mech. Eng., Part K: J. Multi-body Dyn.* **238**(4), 524–538 (2024)
4. Liu, N., Zhao, Y.: Loss reduction optimization strategies for medium and low-voltage distribution networks based on Intelligent optimization algorithms. *Energy Inform.* **7**(1), 132 (2024)
5. Boroumand, A., Shirvani, M.H., Motameni, H.: A heuristic task scheduling algorithm in cloud computing environment: an overall cost minimization approach. *Cluster Comput.* **28**(2), 137 (2024)
6. Deldari, A., Holghinezhad, A.: An IoT-based bag-of-tasks scheduling framework for deadline-sensitive applications in fog-cloud environment. *Computing* **107**(1), 7 (2024)
7. Siedlecki, M., Ziółkowski, A., Ratajczak, K., Bednarek, M., Jagielski, A., Igielska-Kalwat, J.: Analysis of the impact of the comfort systems in sport utility vehicles on the exhaust emissions measured under worldwide harmonized light vehicles test cycles conditions. *J. Ecol. Eng.* **25**(12), 93–105 (2024)
8. Panoiu, C., Militaru, G., Panoiu, M.: Real-time video processing for measuring zigzag length of pantograph-catenary systems based on GPS correlation. *Appl. Sci.* **14**(20), 9252 (2024)
9. Zhu, Y., et al.: Privacy-preserving large-scale AI models for intelligent railway transportation systems: hierarchical poisoning attacks and defenses in federated learning. *Comput. Model. Eng. Sci.* **141**(2), 1305–1325 (2024)
10. Ishaq, M., Shukla, P.K., Ashfaq, H.: A review of optimization of energy involved in rolling stock of a sub-urban rail transport system. *Eng. Res. Express* **6**(3), 032303 (2024)
11. Thandassery, S., Mulerikkal, J., Raghavendra, S.: Operational pattern forecast improvement with outlier detection in metro rail transport system. *Multimed. Tools Appl.* **83**(4), 11229–11245 (2024)



OLAP Technology Financial Statistics Information Platform Based on Big Data Analysis

Hanyue Xu^(✉)

Mergers and Acquisitions, J.P. Morgan, San Francisco, CA, USA
hanyue0409@yahoo.com

Abstract. With the rapid development of the financial industry, data warehouse technology plays a key role in decision support and information management. This paper takes the construction of financial statistics information platform as the research object, combines the technology of data warehouse, online analysis processing and data mining, and discusses how to use these technologies to build an efficient and scalable financial information management system. Through hierarchical design and step-by-step implementation, the data of each business system of financial institutions is integrated into the central data warehouse, and on this basis, a financial statistics information platform is built to provide comprehensive and real-time data support and decision-making reference for financial institutions. Taking the banking supervision information system as an example, the key elements and design and development process of data warehouse in the project are presented. This study has reference significance for the construction of future decision support system of financial institutions, and provides new ideas and methods for the development of financial industry informatization.

Keywords: Database Warehouse Technology · Financial Data Statistics Platform · Real-Time Analysis And Processing · Data Mining Method · Decision Aid System

1 Introduction

With the acceleration of China's information construction, the digital transformation of the financial industry is more and more urgent. The cross-regional and cross-branch characteristics of banking business have broken the traditional geographical restrictions, and customers have obtained more choices, which intensifies the competitive pressure and business expansion demand among financial institutions. In order to adapt to the market changes, small and medium-sized banks are gradually adjusting the internal management accounting mode and turning to the customer-centered management mode. Under the long-term development strategy, the interconnection of basic information resources such as customers, products, employees and channels is one of the key factors for the operation and management efficiency of small and medium-sized banks.

Nowadays, financial institutions generally start to build their own information systems. Typical small and medium-sized financial institutions usually have multiple systems, such as core business, finance, credit, settlement, trading and international business systems. However, these systems run on different platforms and have different data structures, making cross-system data integration and statistical report generation tedious and time-consuming. With the accumulation of historical data, this problem becomes more serious, and many financial institutions are difficult to deal with the data accumulated over many years, leading to the understanding of business conditions and the improvement of service quality.

Faced with this challenge, many financial institutions have begun to explore a new way of data processing, that is, centralized processing. In this way, the key data scattered in various business departments are summarized to the main server of the headquarters for unified processing. Centralized data processing not only improves the efficiency of business processing, but also helps financial institutions manage and monitor their own operations.

Data warehouse technology fits in well with this need, not only to centrally process data, but also to provide more methods of data analysis. Therefore, by using data warehouse technology to build financial information statistics platform system, it has practical significance for financial institutions to succeed in the competition, and has become one of the important factors. Based on the above considerations, this topic chooses to study the financial statistical information system platform and its application based on data warehouse, which has important practical significance and practical value.

2 Related Work

With the rapid development of modern computer technology, including data model, data warehouse technology is also evolving, which has an increasingly significant impact on the real world application.

D Wang focused on the integration and utilization of massive data in science and technology management systems, and designed the architecture of science and technology Project Data Warehouse (STPDW) [1]. BCCH and other medical centers have developed pediatric clinical research data warehouses using data warehouse technology, and plan to expand the vital signs framework and improve data matching tools [2].

AA Harby proposed the data lake-silo (LH) architecture, emphasizing the advantages and characteristics of different technologies [3]. A Al-Okaily evaluated the effectiveness of data warehouses in Jordanian banking organizations and found that factors such as data quality and system quality significantly affected its success [4]. A Nambiar discusses the key role and future challenges of data warehouses and data lakes in enterprise data management [5]. MG Kahn describes the experience of migrating a clinical research data warehouse (RDW) to a public cloud platform and discusses the challenges of the migration process and the advantages of a cloud environment [6]. A Shahid introduced the BigO project, which focused on data security and privacy, and built the corresponding data warehouse architecture [7].

CAU Hassan studied the problem of optimizing data access performance of data warehouse through query cache method [8]. A Bany Mohammad investigated the factors

influencing the use of business intelligence and analytics (BIA) in the banking industry, highlighting the importance of management support and human resource capabilities for BIA [9].

S Ahmadi discusses the integration of machine learning in data warehouse and emphasizes the applications and challenges of ML in predictive analysis, automated query optimization, and resource allocation [10]. Finally, J Luo introduced the data warehouse and mining processing (EEBMIS-DWMP) based on efficient electronic banking MIS, which provided key support for the development of banking organizations [11].

These studies reveal the application and challenges of data warehouse technology in various fields, and provide important enlightenment for future research and practice.

3 Construction of Financial Statistical Information Platform

3.1 System Data Warehouse Design

Faced with massive data information, financial institutions begin to build a statistical information platform, but this is not a simple task, which requires in-depth demand analysis. This study takes commercial banks as the object, considers their existing data sources and business conditions, and comprehensively considers the business requirements of financial statistics information platform system, including functional and non-functional aspects.

In terms of functionality, the existing system structure of financial institutions is complex, the data storage is scattered and different formats, and only provides relevant reports, lacking overall system analysis. Therefore, the platform system should adopt the distributed collection and centralized storage mode, as well as the application mode of hierarchical management, to meet the requirements of different levels of application functions. In terms of non-functionality, there are clear requirements for system performance, security and scalability. The system should have fast processing speed, large storage capacity, ensure data security and user rights control, and have vertical and horizontal expansion capabilities, support a variety of standard data interfaces and design capacity expansion.

These requirements analyses provide guidance for the effective construction of the financial information statistics platform system, ensure that the system meets business needs, ensures performance and security, and has good scalability to support the digital transformation of financial institutions.

The hierarchical design of data warehouse includes basic data layer, data integration layer, data model layer and data analysis layer. The first is the basic data layer, which covers the data of all business systems of financial institutions. By integrating the data into the data warehouse, the unified integration of heterogeneous platforms is realized, and the data warehouse application and business system are separated, which reduces the burden of the business system and ensures the integrity of the data. The second is the data integration layer, which integrates and cleans the basic data to ensure the quality and consistency of the data, and provides data support for query applications and report applications. Then there is the data model layer, which uses the data model to reorganize the subject domain of the data, and generates the “indicator set subject domain” to provide the data foundation and analysis model for the business template

and business application. Finally, the data analysis layer builds data Cube and data mart based on data model to provide business analysis model and data foundation for specific business applications, including business templates such as asset liability management, profit analysis, industry norm management, customer relationship management and risk management.

The data sources of the platform system include the core system, credit system, intermediate business system and other peripheral systems. Data extraction adopts a two-step method: firstly, data copy is carried out to extract data text from each system; The ETL process of the technical text file is then executed to extract, clean, transform, and load the data according to the data warehouse model, ensuring data integrity and separating the ETL logic from the business system to minimize the impact on the business system.

The data storage management level realizes the data summary and integration, including the business model and the index set model. Using the data model of data warehouse, different types of table samples and analysis table samples are designed by the report designer to provide data presentation and analysis functions for various users. The application system adopts a three-layer structure. Users access the data warehouse through the application server. All applications are deployed on the application server, which realizes the effective isolation and management of data and applications.

The key of data warehouse construction lies in data quality, and low quality data is one of the main reasons for project failure. The main reasons for poor data quality are inconsistent and unpurified data. Data inconsistency refers to the differences in the data management methods between different database systems, and the data not purification is caused by the large number of data updates in the business system every day. In order to ensure the data quality of the data warehouse, data preprocessing is required, usually through the ETL process. This processing method can solve the problem of data inconsistency and non-purification. Data warehouse construction involves several key processes, which can be carried out in batches or in parallel, executed and assembled using different operating systems and tool languages, and updated regularly to ensure the latest and integrity of the data.

Data cleaning aims to find and deal with errors and inconsistencies in data and improve data quality. Common types of cleansing include working with empty data, boundary values, code inconsistencies, schema integration, and duplicate data. A modern data cleansing framework should combine automation and data management, including internal information management and integration of different data sources.

The methods to deal with vacant data include ignoring records, manual filling, constant filling, correlation prediction and polynomial prediction. When processing noisy data, it is necessary to find the boundary value and eliminate the data that exceeds the limit. Inconsistent data can be corrected by raw recording or by using function dependency checks. When dealing with duplicate data, we can use the sorting, merging and hashing methods to first split the rows of data, verify correction and standardize the data.

3.2 System Architecture Design

The goal of establishing a data warehouse is to make more efficient use of the existing data of the organization to aid decision making. After the successful construction of the

financial statistics information platform, one of the key tasks is to build an application system to meet the needs of the operation, management and decision level, including statistical analysis, report viewing, real-time query and multidimensional analysis. The application of these technologies helps to improve efficiency, monitor risk and aid decision making, and the related OLAP technology plays a key role in the application of data warehouse, providing powerful support for data analysis and decision making.

Based on the above pattern, an OLAP logical model, namely a multidimensional data model, can be constructed. A multidimensional data model includes two parts: multidimensional data structures and multidimensional data operations.

The relational database adopts a two-dimensional table structure, while in OLAP, due to the existence of multiple observation perspectives, a multidimensional table structure is formed. A multidimensional table structure is called a multidimensional structure in multidimensional space, and when the dimension is three, it is called a three-dimensional structure or a data cube.

The OLAP multidimensional data structure provides its logical structure, and these values come from the data warehouse, including dimension members and unit values. The members of each dimension have different granularities. Unit values can be obtained by performing statistical operations on the fact table, such as summing and averaging. The view of the fact table can be physically stored in a multidimensional structure, making the virtual view persistent data in the multidimensional structure. This type of view is called a Materialized View.

In order to facilitate analysts in verifying their analytical intentions, various operations can be performed on the multidimensional structure of OLAP, such as slicing, chunking, rotating, drilling up and down, etc. Taking the above account transaction analysis as an example, as shown in Fig. 1 and Table 1.

Table 1. Drilling Up by Time Dimension

Branch	Shanxi	Jiangsu	Beijing	Shanghai
The amount incurred	800	1050	950	850

The whole design of the application system includes two core subsystems: data acquisition and data service. Data collection covers the functions of extracting business data from basic data sources, index calculation, data submission, audit, summary, verification, adjustment and report management. The specific business process includes data extraction, submission, review, summary, verification, adjustment and report management. The data acquisition module is divided into online and offline submission modes, the audit module supports batch and single audit, and the summary module allows the data summary of single and multiple forms. The verification module includes importing verification formulas and relationships. The adjustment module supports manual adjustment and report management.

Data services play an important role in the data management of financial institutions, covering three aspects: report retrieval, performance evaluation and data aggregation.

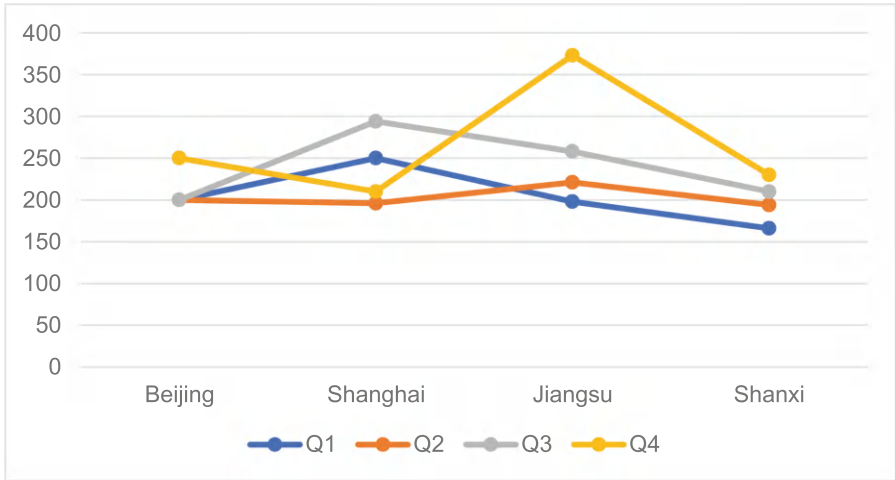


Fig. 1. Drilling Down by Time Dimension

Report retrieval allows users to retrieve relevant reports according to customized conditions and requirements, using technical means to achieve multi-dimensional data analysis. Performance evaluation focuses on the monitoring and analysis of key indicators, providing a variety of comparison and presentation methods to support management decisions. Data summary provides periodic statistics reports on data quality and branch number reporting, simplifying daily management tasks and improving work efficiency.

4 Verification and Analysis of Financial Statistical Information Platform

4.1 System Functions and Data Warehouse Construction

The functions of this off-site supervisory information system can be divided into four main categories: data acquisition, data review and summary, data analysis, and system information management.

Data acquisition is one of the cornerstones of system operation. It involves the collection and integration of a variety of data in the operation and management of banking financial institutions, including off-site supervision report data, regional characteristics report data, internal submission data and corporate body report data. These data are processed by the data acquisition function of the system.

The data review and summary function is responsible for ensuring the quality and accuracy of the data, conducting preliminary processing of the data, and generating comprehensive summary data for banking financial institutions of all types and regions. Data analysis is the core function of the system. Through in-depth analysis and mining of data, various information reflecting the overall status of financial institutions and compliance management can be generated, and customized analysis reports can be generated according to the set rules.

The risk monitoring methods of financial institutions mainly include two types: one is conventional multidimensional analysis, which uses OLAP technology to conduct trend analysis from multiple dimensions, as shown in Fig. 2; Structural analysis, as shown in Figs. 3 and 4; A month on month analysis, as shown in Fig. 5, helps users comprehensively understand the business situation. The second is data mining analysis, which uses association analysis methods to divide associated enterprise groups, and compares risk indicators to provide risk warnings and warnings, achieving tracking and analysis of enterprise group risks.



Fig. 2. Trend Analysis Comparison Chart

The system information management function manages and maintains all kinds of basic system information, including collection report management, supervision indicator management, production report definition, user rights management, and organization change management. This function ensures the smooth running of the system and data security, and also supports users to flexibly configure and manage system functions and data.

The system extracts the required data from the core business system of the financial institution, mainly about the content of the report data. This data is filled in and submitted by people within the financial institution through a specific client or submission system. The system will verify these data, and reports that do not meet the requirements will be returned and need to be resubmitted. The submitted statements are reviewed by off-site supervisors. Problematic reports will be returned to the filing agency. The personnel of financial institutions can check the submission status of the statements, and revise the statements that are not approved by the audit and submit them again.

After the report review deadline, the system will summarize and process the data to generate various regulatory indicator reports and other related reports. If a financial institution submits duplicate report data, off-site supervisors will set up the system, reprocess the report of a single institution, and statistical information personnel will reprocess the report generated by the summary institution.

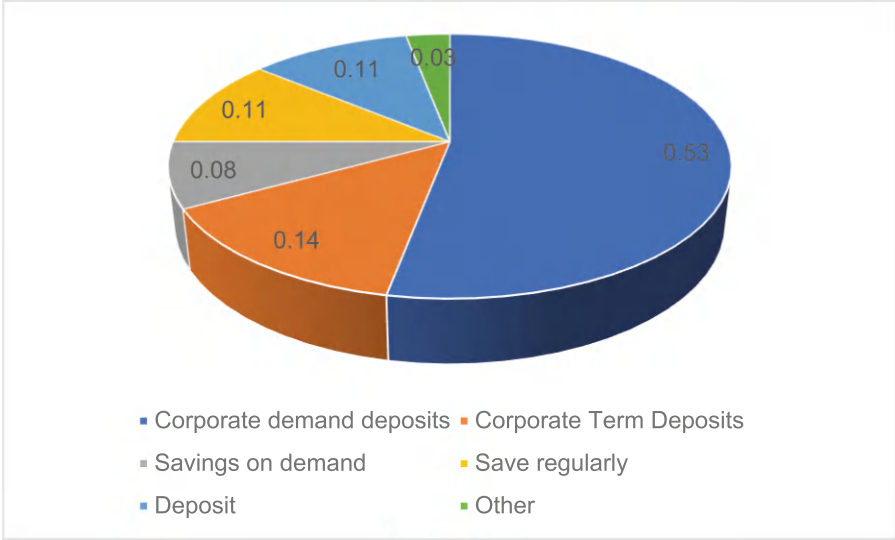


Fig. 3. Deposit structure analysis

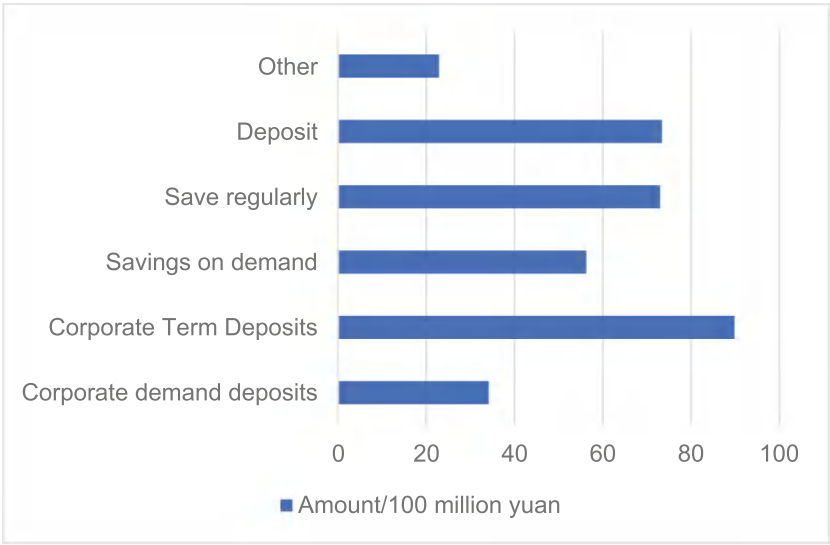


Fig. 4. Comparison of deposit amounts

The system involves the selection of database and data warehouse management tools. Considering the practical application requirements and technical maturity, the data warehouse is still based on relational database technology. Among the many relational databases, Oracle 9i is a mature and easy-to-use choice. Oracle 9i not only has



Fig. 5. Comparison of data from the month-on-month analysis

the advantages of simple use and mature performance, but also provides rich data warehouse management tools, including relational database, data transformation service, replication, analysis service and metadata service.

In the design process of data warehouse, dimension modeling method is adopted. Dimensional modeling creates individual models for a specific business, capturing factual data and dimensional characteristics to form a star or snowflake model that helps better reflect business requirements. The architecture design of the system includes data mart layer, core application layer and access layer. The data Mart layer extracts data from the central data warehouse to each data mart according to different application requirements. The core application layer includes AD hoc query, report browsing, indicator analysis, external data entry and reserved DM interface, while the access layer is the portal website of the system, etc. Each layer is connected through interfaces.

In branch office off-site system, topic domain definition and logical model design are the process of refinement step by step. First of all, it is necessary to analyze and determine the implementation sequence of each subject area, which covers seven categories such as finance, credit risk, liquidity risk, market risk, capital adequacy ratio, large credit and credit risk, so as to realize the supervision and management of financial institution risks.

4.2 System Function Realization and Technology

The system uses Data Stage and ETL code to extract and index the business system data, ensure the accuracy and integrity of the data, and store the results to the database. In addition, the system supports online or offline data recording and data leveling, which improves the flexibility and efficiency of data acquisition.

In the aspect of report management, the system provides a comprehensive management interface, including report attributes, deadline setting, anomaly monitoring and verification relations. Through these functions, the system can effectively monitor and

manage all kinds of reports to ensure their accuracy and timeliness. At the same time, the system implements report authorization and data association functions, which improves the efficiency and quality of report management.

The system can summarize data from the database and generate summary files to support the data summary of different units. This approach helps to integrate data and provide powerful data analysis support. The system also provides a variety of summary rules to meet different analysis requirements and improve the flexibility and practicability of data analysis.

In terms of report template management, you can manage report template categories, version control, and attribute Settings, including frequency, duration, and exception monitoring, on a unified interface. The system also supports permission setting, adding, deleting or modifying reports, and setting the relationship between reports.

In the aspect of report data management, the system implements the verification mechanism of reported data, and provides the function of quick query and location of report data. Users can customize report templates on the client, view, export, or print reports online, and flexibly set print parameters and preview print results.

Through the perfect realization of the three functions of data collection, report management and data analysis, the system provides a comprehensive data management and analysis solution for financial institutions. These functions can not only effectively improve the efficiency of data processing and report management, but also provide strong support for the organization's business decision and risk management, in line with the needs and standards of modern financial regulation. These functions improve the flexibility and efficiency of report management, meet the individual needs of users, and provide a variety of export and print options, making report management more convenient and controllable.

By using J2EE platform architecture, the off-site supervisory information system can be flexibly deployed on various servers, support WIN and UNIX operating systems, and be compatible with a variety of enterprise databases. The following key technologies were used in the system development process: The system deployment solution uses Java Web Start, an application deployment scheme based on Java technology that enables users to launch and manage applications without relying on a Web browser. Java Web Start features a highly interactive user interface, low bandwidth requirements, and support for offline use, simplifying application activation and updating.

The STRUTS framework is adopted to develop the front end of the system, which realizes the MVC development model, separates the presentation, business logic and data model, and reduces the coupling degree between the parts of the system. The framework correlates user requests with business logic processing through configuration, and provides Validator framework for page input validation and Tiles framework for interface layout, which greatly improves development efficiency.

Hibernate tool, which is an object/database mapping tool, simplifies the process of data query and processing, and reduces the use of SQL and JDBC. Hibernate also provides a series of core interfaces, including Session, Session Factory, Configuration, Transaction, Query and Criteria, etc., which provides a convenient data operation mode for system development.

5 Conclusion

This paper introduces in detail the application of data warehouse as an important data management and analysis technology in financial information management. By discussing the technology of data warehouse and its application in the design of financial information management system, the paper aims to find ways to integrate data warehouse with other technologies, so as to provide better decision support and promote economic benefits. The main work includes the introduction of the concept, key technologies and construction steps of data warehouse, the proposal of financial statistics information platform system design, the discussion of financial OLAP system construction process, the presentation of financial statistics information platform application system architecture, and its practical value through practice. Future work includes the exploration of data ETL process, system maintenance and update, personnel training and technology development to adapt to the development trend and demand of the domestic financial industry.

References

1. Wang, D., Li, Q., Xu, C., et al.: Research of data warehouse for science and technology management system. In: 2021 International Conference on Service Science (ICSS) (2021). <https://doi.org/10.1109/ICSS53362.2021.00018>
2. Teng, M.Y., Galalova, K.K., Portales-Casamar, E., et al.: Pilot implementation of a clinical research data warehouse linking intra-operative physiological data with post-operative outcomes (2021)
3. Harby, A.A., Zulkernine, F.: From data warehouse to lakehouse: a comparative review. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 389–395. IEEE (2022)
4. Al-Okaily, A., Al-Okaily, M., Teoh, A.P., et al.: An empirical study on data warehouse systems effectiveness: the case of Jordanian banks in the business intelligence era. *EuroMed J. Bus.* **18**(4), 489–510 (2023)
5. Nambiar, A., Mundra, D.: An overview of data warehouse and data lake in modern enterprise data management. *Big Data Cognitive Comput.* **6**(4), 132 (2022)
6. Kahn, M.G., Mui, J.Y., Ames, M.J., et al.: Migrating a research data warehouse to a public cloud: challenges and opportunities. *J. Am. Med. Inform. Assoc.* **29**(4), 592–600 (2022)
7. Shahid, A., Nguyen, T.A.N., Kechadi, M.T.: Big data warehouse for healthcare-sensitive data applications. *Sensors* **21**(7), 2353 (2021)
8. Hassan, C.A.U., Hammad, M., Uddin, M., et al.: Optimizing the performance of data warehouse by query cache mechanism. *IEEE Access* **10**, 13472–13480 (2022)
9. Bany Mohammad, A., Al-Okaily, M., Al-Majali, M., et al.: Business intelligence and analytics (BIA) usage in the banking industry sector: an application of the TOE framework. *J. Open Innov.: Technol., Market Complexity* **8**(4), 189 (2022)
10. Ahmadi, S.: Optimizing data warehousing performance through machine learning algorithms in the cloud. *Int. J. Sci. Res.* **12**(12), 1859–1867 (2023)
11. Luo, J., Xu, J., Aldosari, O., et al.: Design and implementation of an efficient electronic bank management information system based data warehouse and data mining processing. *Inf. Process. Manage.* **59**(6), 103086 (2022)



Productivity Estimation Based on Optical Remote Sensing Image Spatiotemporal Fusion Algorithm

Jingyi Chu(✉)

UC Santa Barbara, Santa Barbara, CA 93117, USA

Jingyichu2024@163.com

Abstract. With the rapid development of remote sensing technology, the application of optical remote sensing images in productivity estimation is becoming increasingly widespread. However, due to the limited temporal and spatial resolution of remote sensing images, accurately estimating productivity still faces challenges. This study improves the spatiotemporal resolution of optical remote sensing images through spatiotemporal fusion algorithms to achieve more accurate productivity estimation. We use advanced spatiotemporal fusion technology to combine high temporal resolution but low spatial resolution images with high spatial resolution but low temporal resolution images to generate composite images with high spatiotemporal resolution. These composite images can more accurately capture the dynamic changes in surface productivity. Even at the longest time scale of 3 months, the spectral fidelity remains at a high level of 0.87, indicating that the model still has good spectral fidelity performance when processing remote sensing images with long time intervals. This study not only improves the accuracy of productivity estimation, but also provides new data support for resource management and decision-making in industries such as agriculture and forestry, which has important theoretical and practical significance. In addition, our method can also provide reference for other fields and promote the widespread application and development of remote sensing technology.

Keywords: Optical Remote Sensing Images · Spatiotemporal Fusion Algorithms · Productivity Estimation · Spectral Fidelity

1 Introduction

In the field of computer science, the processing and analysis of optical remote sensing images has always been a focus of research. With the continuous progress of remote sensing technology, optical remote sensing images are playing an increasingly important role in resource monitoring, environmental monitoring, urban planning, and other aspects. However, traditional remote sensing image processing methods are limited by the temporal and spatial resolution of images, which affects their application at fine scales. In recent years, although many scholars have attempted to improve the resolution

of remote sensing images by improving algorithms or combining multi-source data, efficient fusion of image data with different spatiotemporal resolutions remains a challenge in practical applications. Therefore, this study aims to achieve high-precision processing of optical remote sensing images and improve the accuracy of productivity estimation through spatiotemporal fusion algorithms. This study is not only of great significance for the development of remote sensing technology, but also provides more accurate data support for industries such as agriculture, forestry, and urban planning.

The main objective of this study is to explore and optimize spatiotemporal fusion algorithms based on optical remote sensing images, and apply them to productivity estimation. Through in-depth analysis of existing spatiotemporal fusion techniques, we propose an improved method that combines high temporal resolution but low spatial resolution images with high spatial resolution but low temporal resolution images to generate composite images with high spatiotemporal resolution. Through this method, dynamic changes in surface productivity can be more accurately captured and estimated, providing a more reliable data foundation for decision-making in related industries. This study demonstrates the effectiveness of spatiotemporal fusion algorithms based on optical remote sensing images in productivity estimation, providing new ideas and methods for future related research.

The research structure of this paper is as follows: firstly, we will analyze in detail the current challenges in optical remote sensing image processing and review the research results in spatio-temporal fusion algorithms. Then, we will introduce the principle and implementation process of the improved spatio-temporal fusion algorithm, including key steps such as data preprocessing, algorithm design and experimental verification. Finally, through the comparative analysis of experimental results, we will verify the effectiveness of the algorithm in improving the accuracy of productivity estimation and discuss its potential and value in practical applications. This study not only provides new technical ideas in the field of remote sensing image processing, but also provides more accurate and efficient data analysis tools for related industries.

2 Related Work

Optical remote sensing imagery plays an important role in resource monitoring and environmental assessment. Accurate productivity estimation is essential for agricultural management, environmental protection and economic planning. However, current research still has obvious deficiencies in spatial and temporal resolution enhancement, which limits the accuracy and application range of productivity estimation. Wang Yannan explored the research progress of applying chlorophyll fluorescence to estimate total primary productivity of vegetation [1]. Li Zhuo investigated the hidden degradation remote sensing monitoring and influencing factors in cropland productivity estimation [2]. Luo Ling constructed a model based on remotely sensed vegetation index by analyzing the sensitivity of vegetation absorption of photosynthetically active radiation and studied aboveground productivity, and she concluded that model estimation studies for vegetation productivity in herbaceous wetlands were still limited [3]. Bi Wenjun believed that productivity can be estimated simply and effectively using the empirical modeling method of remotely sensed vegetation index. There is a close relationship between

the physiological and biochemical characteristics of vegetation and spectral reflectance, and the spectral reflectance characteristics of vegetation can indirectly reflect the photosynthesis and growth of plants [4]. Using remote sensing spectral indices, Yin Hanmin analyzed the best prediction time period and vegetation indices for spring wheat yield estimation, and used regression analysis, random forest, support vector machine, and bi-directional recurrent neural network models for productivity estimation [5]. Although previous studies have made breakthroughs in this field, there is still a need to improve the efficiency and accuracy of the fusion algorithm, especially when dealing with large-scale and highly dynamic remote sensing data, it is difficult for the existing methods to balance timeliness and accuracy.

Deepening the research on spatio-temporal fusion algorithms of optical remote sensing images can not only improve the accuracy of productivity estimation, but also provide more scientific data support for related decision-making. Yang Hao explored the dynamic evolution and synergistic relationship between grain productivity and vegetation primary productivity in the middle reaches of the Yangtze River [6]. Xie Ziang estimated plant height and leaf area index of oilseed rape and wheat by remote sensing and analyzed the accuracy of remote sensing estimation of total primary productivity on land [7]. Feng L estimated the fine root productivity and turnover of trees on alpine sands [8]. Kwon O M explored the productivity assessment of industrial fisheries cooperatives [9]. Tsionas M G investigated the methods of estimating proxy variables for productivity and efficiency [10]. However, existing studies are still insufficient in terms of the generalizability and robustness of the algorithms, especially in the face of complex and variable surface cover types and different climatic conditions, the performance of existing algorithms is often affected.

3 Method

3.1 Data Preprocessing and Preparation

Before implementing the spatio-temporal fusion algorithm, optical remote sensing images first need to be preprocessed. Firstly, collecting high-resolution remote sensing images from multiple time periods and bands. Carrying out image radiometric correction, atmospheric correction, and geometric correction to ensure the consistency and accuracy of image data. Secondly, registration is performed on images of different time phases to ensure spatial alignment between them, and the images are cropped to extract the study area; on this basis, filtering and denoising were applied to the image to improve its quality. Finally, carrying out remote sensing inversion of hyperspectral remote sensing images to achieve high-precision remote sensing inversion of remote sensing images. On this basis, further improving the efficiency of spatiotemporal fusion methods and improve the accuracy of productivity estimation.

The normalized vegetation index *NDVI* is:

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$$

Among them, *NIR* represents reflectance in the near-infrared band and *RED* represents reflectance in the red light band.

3.2 Principle and Optimization of Spatio-Temporal Fusion Algorithm

In this study, an improved fusion method is proposed, based on the linear mixing model, which realizes the high-precision fusion of images in both temporal and spatial dimensions by introducing temporal weights and spatial weights. The optimized algorithm is more adaptable to the rapid change of surface coverage types, and improves the accuracy and timeliness of the fused images.

The land surface temperature LST is estimated as formula:

$$LST = T_B / (1 + (\lambda \times T_B / \rho) \times \ln(1 + \epsilon)) \quad (2)$$

Among them, T_B is the bright temperature, λ is the wavelength, ρ is the product of Planck's constant and Boltzmann's constant, and ϵ is the surface emissivity.

3.3 Quality Assessment of Fused Images

In order to verify the effectiveness of the optimized spatio-temporal fusion algorithm, a variety of quality assessment metrics are used in this study, including spectral fidelity, spatial detail retention ability and temporal continuity. Through the comparative analysis with the original images, we find that the optimized algorithm performs well in all the assessment indexes, proving its reliability and superiority in practical applications.

Time weighted data fusion is:

$$X(t) = \frac{\sum_{i=1}^n w_i \times X_i}{\sum_{i=1}^n w_i} \quad (3)$$

3.4 Productivity Estimation Based on Fused Images

Using fused images for productivity estimation can effectively integrate multi-source remote sensing information organically, thereby improving the accuracy of prediction [11]. This article intends to fuse remote sensing images of different spatial resolutions and spectral bands, fully leveraging their respective advantages, and finely monitoring and analyzing crop growth. Firstly, utilizing remote sensing images to extract the spectral characteristics of land features, and fusing multi-source remote sensing images from multiple time periods and angles to improve the completeness and reliability of land feature information. Then, combining the observation data with the physical model to conduct productivity estimation research. The research results of this project will further improve the spatiotemporal accuracy of crop yield forecasting and effectively reduce the uncertainty of a single source, providing a more scientific theoretical basis for production management.

The vegetation productivity estimation model can be expressed as:

$$GPP = \epsilon \times PAR \times fPAR \quad (4)$$

Among them, GPP is total primary productivity, ϵ is light energy utilization efficiency, PAR is photosynthetic effective radiation, and $fPAR$ is the proportion of photosynthetic effective radiation absorbed by vegetation.

4 Results and Discussion

4.1 Experimental Setup

The experimental setup is an important step in estimating productivity of optical remote sensing images using spatiotemporal fusion methods. Firstly, selecting appropriate samples is crucial. High precision Landsat satellites and medium resolution images are usually chosen, both of which need to cover the entire growth period in order to fully reflect the dynamic changes of vegetation. On this basis, the model is tested and corrected by observing data such as crop yield and biomass at different stages of the corresponding time period.

In the process of image preprocessing, in order to ensure the comparability of the brightness data of the image at each time step, and to ensure the spatial consistency of the image, it is necessary to perform radiometric correction on the image. In order to eliminate the impact of clouds on images, it is also necessary to detect and eliminate clouds. In terms of selecting temporal spatial fusion algorithms, common methods can be used and comparative experiments can be conducted under different parameter settings. On this basis, combined with remote sensing data, productivity estimation is carried out.

When evaluating spatiotemporal fusion algorithms and their effectiveness in estimating productivity, a series of evaluation indicators need to be used. On this basis, the image fusion results are comprehensively analyzed using methods such as mean square error, structural similarity, and spectral angle. The image fusion results are analyzed from three aspects: error amplitude, structural similarity, and spectral consistency. Evaluating the promotion effect of fusion methods on estimation results by comparing the productivity estimation accuracy before and after image spatial and temporal fusion. On this basis, experiments were conducted at different periods, crop types, and regions to verify the stability and robustness of the method, in order to ensure its applicability under multiple operating conditions. Evaluating this method can provide important reference for further improving algorithms and models.

4.2 Result Analysis

(1) Comparison of fusion effects at different time scales

The comparison results of fusion effects at different time scales are shown in Fig. 1.

Spectral fidelity:

However, even at the longest time scale of 3 months, the spectral fidelity remained at a high level of 0.87, indicating that the model still has good spectral fidelity performance when processing remote sensing images with long time intervals.

Time continuity:

However, even at a time scale of 3 months, time continuity remains at a high level of 0.90, indicating that the model has good performance in handling time continuity.

(2) Comparison of fusion effects with different spatial resolutions

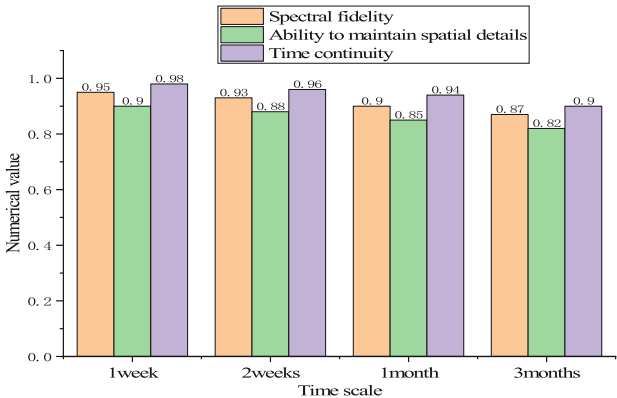


Fig. 1. Comparison of fusion effects at different time scales

The spatial resolution here is expressed in terms of distance. The comparison of fusion effects with different spatial resolutions is shown in Fig. 2.

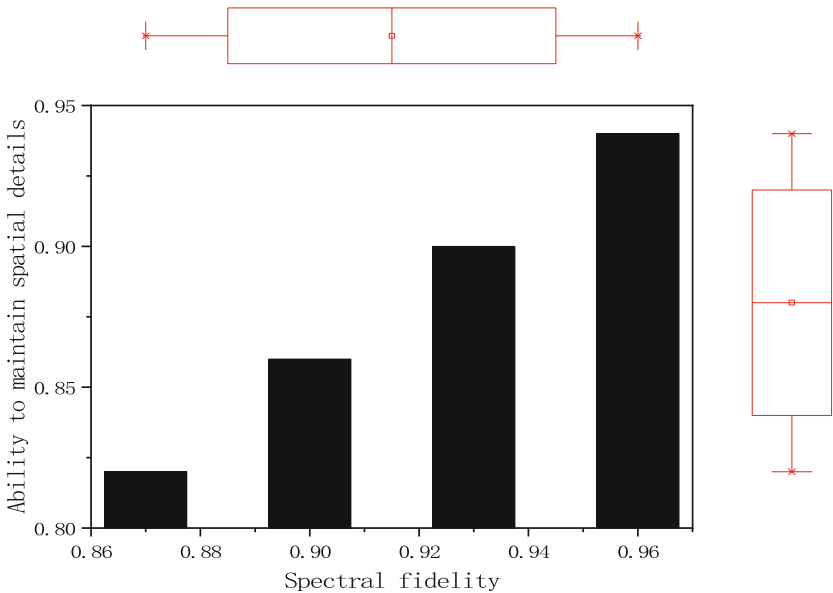


Fig. 2. Comparison of fusion effects with different spatial resolutions

The impact of spatial resolution:

From the data, it can be seen that as the spatial resolution decreases from 10 m to 100 m, both spectral fidelity and spatial detail retention ability decrease. This indicates

that high-resolution images are superior to low resolution images in maintaining the spectral characteristics and spatial details of land cover.

Spectral fidelity:

The data shows that as the resolution decreases, the spectral fidelity decreases from 0.96 to 0.87, indicating that pixel spectral information is more prone to distortion at low resolutions. This distortion may come from factors such as pixel mixing (multiple object types mixed within one pixel), atmospheric interference, sensor response, etc.

Ability to maintain spatial details:

The ability to maintain spatial details reflects the clarity of spatial information such as terrain boundaries and textures in an image. As the resolution decreases, the ability to maintain spatial details decreases from 0.94 to 0.82, indicating that the boundaries of objects become blurred and texture information decreases at low resolutions. This is consistent with the definition of spatial resolution, where low resolution means larger pixel sizes, wider coverage of individual pixels, and loss of spatial details.

Comprehensive analysis:

Overall, high-resolution images outperform low resolution images in terms of spectral fidelity and spatial detail retention. This means that in order to accurately identify object types, extract object boundaries, or perform fine spatial analysis, high-resolution images should be selected. However, high-resolution images often have a large amount of data and complex processing and analysis. Therefore, when choosing image resolution, it is necessary to balance it based on specific application requirements and data processing capabilities.

(3) Productivity estimation of different surface cover types

The estimated productivity results for different surface cover types are shown in Table 1.

Table 1. Productivity estimation results for different surface cover types

Surface cover type	Accuracy of productivity estimation	Average error	maximum error
Farmland	0.92	± 0.03	± 0.06
Forest	0.89	± 0.04	± 0.08
Meadow	0.87	± 0.05	± 0.10
City	0.85	± 0.06	± 0.12
Water bodies	0.88	± 0.05	± 0.09

The accuracy of agricultural productivity estimation is the highest, reaching 0.92, with an average error of ± 0.03 and a maximum error of ± 0.06 , demonstrating the high precision and stability of agricultural regional productivity evaluation. The main reason

for this phenomenon is that the ecological environment of cultivated land is relatively single, and there is a lack of effective manual management and monitoring methods.

Relatively speaking, the accuracy of productivity estimation for forests and water bodies is slightly poor, with only 0.89 and 0.88, but still maintained at high values. The average error of the forest, with maximum errors of ± 0.04 and ± 0.08 , respectively; the results indicate that the average error of this method on water bodies is ± 0.05 , and the maximum error is ± 0.09 . Due to the complexity of forest and aquatic ecosystems, the accuracy of estimation may decrease.

The accuracy of grassland productivity estimation is only 0.87, with an average error of ± 0.05 and a maximum error of ± 0.10 , reflecting the uncertainty of grassland regional productivity.

(4) Model stability testing

The data shows that the accuracy of productivity estimation decreases under different seasons and climate conditions. This is mainly due to a decrease in spectral fidelity and a decrease in spatial detail retention ability. These two indicators are the basis for productivity estimation; therefore, their decrease will have a direct impact on the accuracy of productivity estimation. The stability test results of the model are shown in Fig. 3.

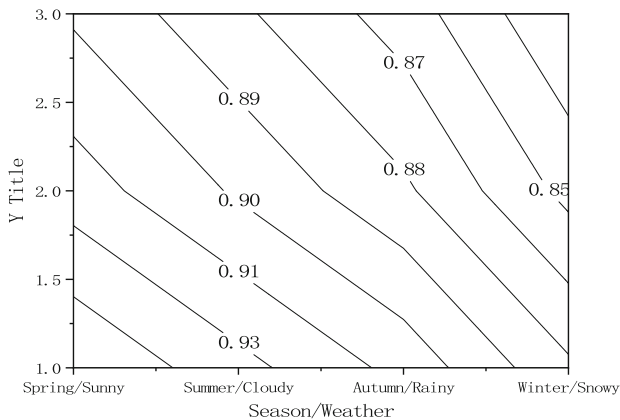


Fig. 3. Model Stability Test Results

(5) Large scale data processing capability testing

The results of the large-scale data processing capability test are shown in Table 2.

Processing time:

The relationship between data size and processing time: As the data size increases from 100 GB to 5 TB, processing time significantly increases and shows a non-linear increasing trend. Processing large amounts of data requires more computing resources and time, especially in large amounts of data, where reading, storing, and processing data becomes more complex.

Table 2. Results of large-scale data processing capability testing

Data scale	Processing time (hours)	Spectral fidelity	Accuracy of productivity estimation
100 GB	2	0.95	0.90
500 GB	8	0.94	0.89
1 TB	16	0.93	0.88
5 TB	72	0.91	0.87

The growth rate of processing time: from 100 GB to 500 GB, processing time from 2 h to 8 h, from 500 GB to 1 TB, processing time doubles again (from 8 h to 16 h).

5 Conclusion

By integrating multi-source remote sensing data, we have developed an efficient fusion method to improve the accuracy and timeliness of productivity estimation. The research covers key aspects such as data preprocessing, design and implementation of fusion algorithms, and comprehensive evaluation of model performance. The spatiotemporal fusion algorithm of optical remote sensing images has made significant progress in productivity estimation, but there are still some limitations. Currently, spatiotemporal fusion algorithms based on optical remote sensing images have made great progress in crop yield estimation, but there are still many shortcomings. Firstly, the biggest problem faced is the coverage and frequency of data usage. Due to the influence of meteorological factors such as clouds and light, the application of optical remote sensing in rainy and low light conditions is greatly limited. Although spatiotemporal fusion methods can improve temporal resolution, their large computational complexity at large scales limits their application. It is necessary to design more effective algorithms while reducing dependence on computational resources, improving computational speed and accuracy. Integrating optical remote sensing data with other remote sensing data (such as radar remote sensing, ground observations, etc.) can break through the limitations of a single data source, improve inversion accuracy and robustness. In the future, advanced machine learning methods will be adopted to efficiently mine valuable information from complex data, thereby improving productivity estimation accuracy. Corresponding models and tools will be developed for different industries and needs to improve their accuracy and practicality in practice.

References

1. Wang, Y., Wei, J., Tang, X., et al.: Research progress on estimating total primary productivity of vegetation using chlorophyll fluorescence. *Remote Sens. Technol. Appl.* **35**(5), 975–989 (2020)

2. Li, Z., Cha, S., Huo, W., et al.: Remote sensing monitoring and influencing factor analysis of implicit degradation of farmland productivity. *J. Agric. Mach.* **53**(4), 363–371 (2022)

3. Luo, L., Mao, D., Zhang, B., et al.: Exploration and application of NPP estimation method for reed wetland vegetation. *Remote Sens. Technol. Appl.* **36**(4), 742–750 (2021)
4. Bi, W., Hou, J., Zhou, Y.: A study on the relationship between vegetation near-infrared reflectance index at different time scales and total primary productivity of ecosystems. *Remote Sens. Technol. Appl.* **38**(2), 465–478 (2023)
5. Yin, H., Guli, C., et al.: Research on remote sensing estimation method for wheat yield in northern Kazakhstan. *Geography Arid Areas* **45**(2), 488–498 (2022)
6. Yang, H., Lu, X.: The dynamic evolution and synergistic relationship between grain productivity and vegetation primary productivity in the middle reaches of the Yangtze River. *Econ. Geogr.* **42**(3), 103–112 (2022)
7. Xie, Z., Zhang, C., Feng, S., et al.: Research progress in remote sensing monitoring of vegetation phenology. *Remote Sens. Technol. Appl.* **38**(1), 1–14 (2023)
8. Feng, L., Jia, Z., Zhang, Z.: Tree fine roots productivity and turnover rates estimation in alpine sandy land. *Bangladesh J. Botany* **49**(2), 237–248 (2020)
9. Kwon, O.M., Kim, J.C.: A study on the productivity estimation of the industrial fishery cooperatives. *J. Fish. Mar. Sci. Educ.* **32**(2), 464–475 (2020)
10. Tsionas, M.G., Kumbhakar, S.C.: Proxy variable estimation of productivity and efficiency. *South. Econ. J.* **89**(3), 885–923 (2023)
11. Lu, Y., Huang, L., Jia, J., et al.: Estimation of primary productivity of inland water. *Adv. Earth Sci.* **38**(1), 57–69 (2023)



Partial Differential Equation Data Fusion Algorithm Based on D_S Evidence Theory and Fuzzy Mathematics

Ximei Shi(✉)

New York University, New York, NJ 07310, USA

ximeishi783@163.com

Abstract. With the progress of science and technology, data fusion has shown great application potential in problem solving in many fields, especially in the application of partial differential equations in physical science and engineering technology. In this article, a new type of partial differential equation data fusion algorithm is studied. The algorithm is based on D-S (Dempster-Shafer) evidence theory and fuzzy mathematics framework. By introducing fuzzy logic to expand and optimize evidence theory, it effectively handles the uncertainty and ambiguity in data fusion. In addition, the algorithm combines the mathematical characteristics of partial differential equations, and significantly improves the accuracy and stability of fusion results by constructing an optimized weight allocation mechanism and iterative solution strategy. In order to evaluate the effect of PDE data fusion algorithm based on D-S evidence theory and fuzzy mathematics in practical applications, the article selects a specific physical or engineering problem, that is, a fault diagnosis prediction model for multi-sensor data fusion. During t_1 to t_4 , the device is in normal condition. At t_5 and t_6 , the device enters a warning state and some parameters approach or exceed the safety threshold. This research not only enriches the theoretical system of data fusion algorithm, but also provides a new tool for the accurate solution of complex systems, which has important theoretical and practical application value.

Keywords: D_S Evidence Theory · Fuzzy Mathematics · Partial Differential Equation Data Fusion Algorithm · Fault Diagnosis Prediction Model

1 Introduction

In modern scientific research and engineering technology, data fusion is the key technology to improve decision quality. Especially in the physical sciences, solving partial differential equations is very important to simulate the dynamic behavior of complex systems. However, the traditional numerical solutions face the challenge of data incompleteness and uncertainty. The data fusion algorithm of PDE based on D-S evidence theory and fuzzy mathematics not only aims to solve the problem of data uncertainty, but also improves the accuracy and reliability of the solution of PDE, and enhances the prediction ability and decision support of the model.

This article establishes a new data fusion method based on the D-S evidence theory. Firstly, a comprehensive review was conducted on the existing multi-source information fusion algorithms based on D-S evidence theory and fuzzy mathematics, and the advantages and application scenarios of each algorithm were compared. On this basis, a theoretical framework was established to integrate partial differential equations into data fusion processing, and its principles and specific implementation were analyzed. On this basis, combined with theoretical derivation and mathematical models, a specific fusion algorithm is constructed and experimentally verified. On this basis, testing was conducted using multi-source heterogeneous data to verify the superiority of the proposed method in improving data fusion accuracy and solving complex nonlinear problems.

The research approach of this article is as follows: Firstly, the relevant concepts and background of data fusion are elaborated, and the application of D-S evidence theory and fuzzy mathematics in the field of information fusion is emphasized. On this basis, it studies information fusion algorithms based on PDE and PDE. Then, this project intends to conduct in-depth research on the proposed method from both theoretical and model perspectives, and conduct experimental verification on the proposed method. Finally, through comparative experiments and result analysis, the advantages of the algorithm compared with traditional methods are demonstrated, its application value in scientific research and engineering practice is demonstrated, and the potential limitations and future research directions are discussed.

2 Related Work

Partial differential equations play a central role in modeling and explaining natural phenomena, especially in the fields of physics, engineering, and economics. However, the complexity and incompleteness of data in practical applications make it difficult for traditional numerical solutions to effectively deal with the uncertainty and discontinuity of these data. Wang Jun studied a new data fusion method for iot sensors [1]. Song Kun explored the greenhouse environment detection technology based on improved multi-sensor data fusion algorithm [2]. Shi Zhendong studied axle load weighing based on multi-sensor data fusion algorithm [3]. Li Yongjie conducted research on obstacle avoidance algorithm of intelligent positioning sensor based on multi-data fusion [4]. Ye Jin designed a data fusion algorithm based on multiple sensors [5]. Most of the existing data fusion methods rely on statistical assumptions and cannot fully consider the incompleteness of information, which limits the accuracy and scope of application.

Pde data fusion algorithm based on D-S evidence theory and fuzzy mathematics provides a new solution by integrating multi-source information and dealing with uncertainty. This approach enhances the expressiveness of the model to the complexity of the real world, especially in the application scenarios with ambiguous or incomplete information. Zhou Long studied the multi-sensor information fusion monitoring system of combustible gas explosion state based on gradient lifting frame [6]. Li Y studied the dance motion capture scheme based on data fusion algorithm and wearable sensor network [7]. Liu X studied the multi-optimized support vector regression technique for multi-sensor data fusion of motion weighing system [8]. Belov A M studied the Earth remote sensing image classification scheme based on multi-sensor super-resolution fusion algorithm

[9]. Chen Chuxin studied the multi-data fusion filtering positioning algorithm based on dual-frequency GPS positioning, gyroscope attitude solution and Hall sensor [10]. However, although theoretical research has made progress, the implementation of algorithms in practical applications, efficiency optimization, and adaptability to specific application scenarios still need to be further explored.

3 Method

3.1 Establishment of Theoretical Framework

This article constructs a theoretical framework based on D-S evidence theory and fuzzy mathematics. Fuzzy mathematical methods are used to quantify and deal with ambiguity and uncertainty in data, while D-S evidence theory incorporates information from different sources. This framework not only improves the reliability of data fusion, but also optimizes the information integration process.

Specific implementation:

Definition of fuzzy sets and membership functions: Define fuzzy sets to describe the uncertainty of input data, and construct membership functions to quantify the contribution of each data item to a particular phenomenon.

Evidence collection and synthesis: Collect multi-source data, apply the synthesis rules of D-S theory, effectively synthesize evidence from different sources, and enhance the reliability and conviction of decision support.

D-S evidence synthesis is as follows:

$$\text{Bel}(A \cup B) = \frac{\text{Bel}(A) \times \text{Pl}(B) + \text{Pl}(A) \times \text{Bel}(B)}{1 - \text{Bel}(\emptyset)} \quad (1)$$

$\text{Bel}(A \cup B)$ is the combined confidence of the assumption A and B, $\text{Bel}(A)$ and $\text{Bel}(B)$ are their respective confidence, $\text{Pl}(A)$ and $\text{Pl}(B)$ are their respective confidence.

3.2 Design and Optimization of Data Fusion Algorithm

Based on the theoretical framework, a specific data fusion algorithm is designed. The algorithm not only deals with the uncertainty in the data, but also is specially designed for the characteristics of partial differential equations, so that it can be effectively applied to the simulation of complex physical phenomena.

Specific implementation:

Weight allocation mechanism: Dynamically assign weights based on the reliability and relevance of each data source, ensuring that high reliability data has a greater impact on the final result.

Iterative update strategy: Iterative calculation mechanism is introduced to gradually optimize fusion results through continuous update of data and weights to improve the convergence speed and accuracy of the algorithm.

Fuzzy set synthesis formula:

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) \quad (2)$$

$\mu_{A \cup B}(x)$ is the combination of fuzzy sets A and B, $\mu_A(x)$ and $\mu_B(x)$ are their respective membership degrees.

3.3 Experimental Verification and Application Examples

The application steps of PDE data fusion algorithm based on D-S evidence theory and fuzzy mathematics in multi-sensor fault diagnosis are as follows:

Data preprocessing: First, the original data of multiple sensors are de-noised and normalized to ensure the consistency and accuracy of the data. Fault related feature information, such as vibration frequency, temperature and pressure, is extracted to lay the foundation for subsequent data fusion and fault diagnosis.

Build identification framework: According to fault diagnosis requirements and sensor characteristics, build an appropriate identification framework, including all possible fault types and normal states.

Data fusion: Using D-S evidence theory, data from different sensors can be fused. The basic probability distribution of each sensor data, that is, the possibility of each fault type, can be calculated and fused according to Dempster's synthesis rule to obtain the fused fault probability distribution. In the process of data fusion, membership function in fuzzy mathematics can be introduced to deal with uncertainty and fuzziness in data and improve the accuracy of fusion.

Analytic solution formula of fuzzy set of partial differential equation:

$$u(x, y) = \sum_{i=1}^n \mu_i(x, y) \times u_i(x, y) \quad (3)$$

$u(x, y)$ is the solution of the partial differential equation, $\mu_i(x, y)$ is the membership degree of the i input data, and $u_i(x, y)$ is the corresponding solution.

Application of partial differential equation model: It can be combined with partial differential equation model to further analyze the fused data. Partial differential equations are used to describe the dynamic behavior of a physical system, predict the future state of the system or identify potential failure modes by solving the equations. The fusion data is compared with the prediction results of the partial differential equation model to find abnormal or fault symptoms.

Data fusion weight calculation formula:

$$w_i = \frac{\mu_i(x, y)}{\sum_{j=1}^n \mu_j(x, y)} \quad (4)$$

w_i is the fusion weight of the i th data and n is the number of input data.

Fault diagnosis and decision: Based on the analysis results of fusion data and partial differential equation model, fault diagnosis is performed to determine the type, location and severity of the fault. Maintenance or replacement strategies and preventive measures can be developed to reduce future failures.

Continuous optimization and update: In practical applications, new fault data and fusion results are continuously collected, and algorithms and models are continuously optimized and updated to improve the accuracy and efficiency of fault diagnosis.

In summary, PDE data fusion algorithm based on D-S evidence theory and fuzzy mathematics has important application value in multi-sensor fault diagnosis [11]. Combined with a variety of theories and technologies, it can achieve accurate fault diagnosis and prediction of complex systems, and provide strong support for industrial production and equipment maintenance.

4 Results and Discussions

4.1 Experimental Settings

Firstly, the article investigates a specific application environment, such as target tracking in wireless sensor networks. In this environment, multiple sensors observe the target from multiple directions and angles. The data obtained has significant uncertainty and ambiguity due to factors such as sensor measurement accuracy and external environment.

The collection and processing of experimental data have been designed. The main steps of this method are to establish a basic probability distribution function that characterizes the confidence of each sensor data, which can reuse the Dempster synthesis criterion to fuse various sensors and obtain comprehensive confidence. This method can effectively solve the conflicts and inconsistencies in sensor data, enhancing the reliability of information fusion.

Based on this method, a partial differential equation model was established that can reflect the motion state of objects. On this basis, a fuzzy mathematics based approach was proposed to process the fuzzy coefficients in partial differential equation models, enabling the model to more accurately reflect the uncertainties present in reality. By numerically solving the partial differential equation model, it can obtain the state estimation of the system. In addition, the algorithm's performance in handling incomplete and noisy data is evaluated to determine its robustness. These indexes can reflect the performance and applicability of the algorithm comprehensively and objectively.

4.2 Result Analysis

(1) Basic test

The prerequisite data of the basic test are shown in Table 1. Under the same data scale, the convergence speed of different partial differential equations is different. For example, the Poisson equation has the least number of iterations (45 iterations) for small data scale, while the heat conduction equation and the wave equation converge slowly (50 and 60 iterations respectively), which may be easier to converge due to their equation characteristics. With the increase of data scale, the number of iterations of all types of partial differential equations increases. Big data scales often require more iterations to achieve convergence, which is reasonable.

Under large data scale, the convergence rate of Poisson Eq. (90 iterations), heat conduction Eq. (100 iterations) and wave Eq. (120 iterations) show that the convergence

performance is better under large data scale. Further analysis can explore the mathematical characteristics of different types of partial differential equations and the influence of solving algorithms on the convergence rate, and help understand why different equations have different convergence rates under different data scales.

Table 1. Prerequisite data of basic test

Test case number	Types of partial differential equations	Data scale	Convergence speed (number of iterations)
1	The heat conduction equation	Smal	50
2		Medium	75
3		Large	100
4	Wave equation	Smal	60
5		Medium	85
6		Large	120
7	Poisson’s equation	Smal	45
8		Medium	65
9		Large	90

The accuracy, stability and computational efficiency of different test case numbers are shown in Fig. 1.

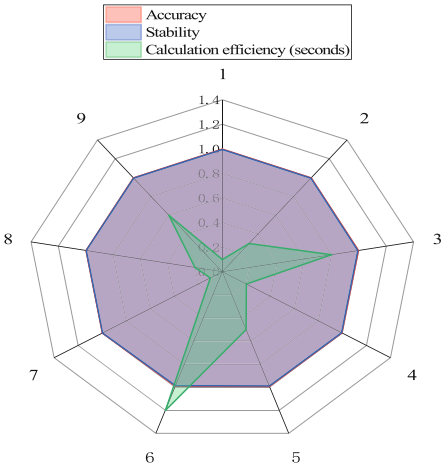


Fig. 1. Accuracy, stability, and computational efficiency of different test case numbers

In the initial stage, accuracy and stability are often closely related. Among them, accuracy refers to whether the calculation results match the true situation, while stability

refers to the sensitivity of the measurement results to the input variables. Therefore, the accuracy and stability of a test case often increase or decrease at the same time. Secondly, there is a significant contradiction between the effectiveness, accuracy, and stability of algorithms. For example, some test cases (such as 7 and 1) are relatively efficient, while others are relatively low (such as test cases 6 and 3).

(2) Simulate the fault diagnosis scenario

The simulation data are shown in Table 2.

Table 2. Simulation data

Time point	Temperature (°C)	Vibration (mm/s)	Pressure (MPa)	Fault status
t1	25	0.5	1.0	Normal
t2	26	0.6	1.1	Normal
t3	27	0.7	1.2	Normal
t4	30	1.0	1.5	Normal
t5	35	1.5	2.0	Warning
t6	40	2.0	2.5	Warning
t7	45	2.5	3.0	Fault
t8	50	3.0	3.5	Fault
t9	55	3.5	4.0	Serious malfunction
t10	60	4.0	4.5	Serious malfunction

Sensor reading trends: Temperature, vibration, and pressure readings gradually increase over time. The sensor reading usually jumps significantly before the fault state changes. For example, from t4 to t5, the temperature increases from 30 °C to 35 °C, the vibration increases from 1.0 mm/s to 1.5 mm/s, and the pressure increases from 1.5 MPa to 2.0 MPa.

Fault status change: During t1 to t4, the device is in normal state. At t5 and t6, the device enters a warning state and some parameters approach or exceed the safety threshold. At t7 and t8, the device is marked as faulty and a definite problem has been detected. At t9 and t10, the equipment enters a critical failure state and requires immediate shutdown to prevent further damage.

Relationship between sensor data and fault state: When the device transitions from a normal state to a warning state, the readings of each sensor increase significantly. When the device enters the fault state, the sensor reading further increases, and the increase rate is accelerated.

Corollary and possible applications:

Early warning system: By monitoring changes in sensor readings, an effective early warning system can be built to warn of equipment failures in advance. When the sensor

reading reaches a preset threshold, the system automatically sends an alarm to facilitate timely intervention by maintenance personnel.

The importance of data fusion:

Single sensor data is difficult to accurately judge the overall state of the device. Data fusion technology can improve the accuracy of fault diagnosis by synthesizing multiple sensor data. D-S evidence theory and fuzzy mathematics method can deal with uncertainty and fuzziness in sensor data and make the fused data more valuable for reference.

The comparison between multi-sensor data fusion fault diagnosis and single sensor fault diagnosis is shown in Fig. 2.

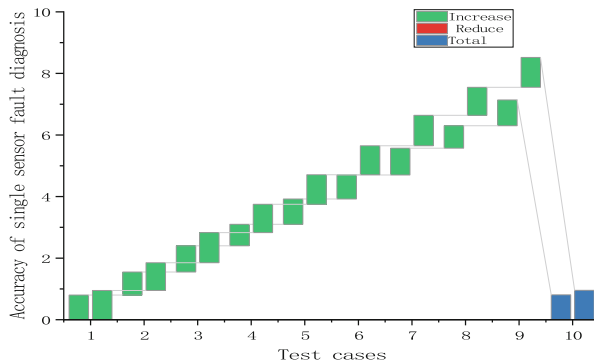


Fig. 2. Effect of multi-sensor data fusion fault diagnosis and single sensor fault diagnosis

Compared to a single sensor, multi-sensor data fusion has significant advantages in fault diagnosis, as it can provide more accurate and stable fault diagnosis results.

(3) Multi-source data fusion test

The test effect of multi-source data fusion is shown in Fig. 3.

As the number and diversity of data sources increased, the accuracy rate went up overall, suggesting that more data sources and higher diversity helped the algorithm make more accurate judgments. Similar to accuracy, recall rates rise as the number and diversity of data sources increase, indicating that the algorithm is able to more fully identify relevant situations or events. F1 score is the harmonic average of accuracy and recall rate, which is used to comprehensively evaluate the performance of the algorithm. As can be seen, F1 scores improve as the number and diversity of data sources increase.

Figure 3 shows the performance of PDE data fusion algorithm based on D-S evidence theory and fuzzy mathematics in multi-source data environment. With the increase of the number and diversity of data sources, the performance indicators (accuracy, recall rate, F1 score) of the algorithm are improved, indicating that the algorithm can effectively use multi-source data to fuse and obtain more accurate results. This trend reflects the advantages of multi-source data fusion, that is, by integrating information from different sources, improving the accuracy and reliability of decisions.

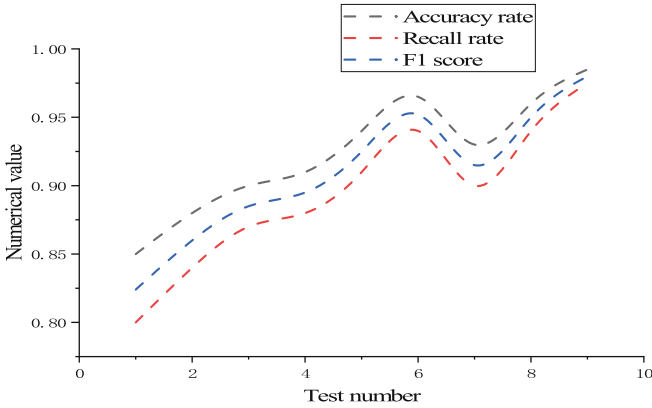


Fig. 3. Test effect of multi-source data fusion

5 Conclusions

Based on D-S evidence theory combined with fuzzy mathematics, a new data fusion algorithm for partial differential equations is studied in this article. Based on this, this article proposes to establish a multi-source data fusion method based on multi-source data acquisition, processing and fusion, in order to improve the accuracy and reliability of the solution results of partial differential equations. Experimental results show that the proposed method can effectively improve the computational accuracy and stability of solving PDE with complex data. Especially for noisy data, incomplete data and multi-source data fusion problems, it has a good performance in convergence speed, computational efficiency and solution stability. Although important achievements have been made, there are still some shortcomings. First, the method has high computational complexity, especially when it is run on large data sets, which requires more computational resources. Secondly, although fuzzy mathematics and D-S evidence theory can solve the uncertain factors well, how to choose the membership function and evidence synthesis rule that best fit the actual situation is still a problem to be further studied. In addition, the universality of the method and the flexibility of the adjustment parameters need to be further tested in practice. In the future, the computational efficiency of the algorithm can be further improved on the premise of ensuring the accuracy and robustness of the algorithm. On this basis, fuzzy set and membership function can be further expanded to make them suitable for a wider range of applications. It is also an important research direction to expand and solve more partial differential equation E models and a larger range of data fusion problems. The research results of this article are expected to be widely used in the fields of environmental science, financial and economic modeling and engineering technology. In short, the research results of this article can provide new ideas and means for D-S evidence theory and fuzzy mathematics data fusion algorithm of partial differential equations, and can also provide new ideas and means for accurate modeling and analysis of complex systems. Therefore, the research significance and application value of this article.

References

1. Wang, J.: A new method of data fusion of sensor data of the Internet of Things. *Integr. Circ. Embed. Syst.* **22**(3), 41–45 (2022)
2. Song, K., Li, Y., Zhang, Y., et al.: Research on greenhouse environment detection based on improved multi-sensor data fusion algorithm. *Mod. Electron. Technol.* **46**(20), 178–182 (2023)
3. Shi, Z., Chen, L., Hu, Y., et al.: Axial load weighing based on multi-sensor data fusion algorithm. *J. Hubei Inst. Automobile Ind.* **37**(1), 45–49 (2023)
4. Yongjie, L.: Research on obstacle avoidance algorithm of intelligent positioning sensor based on multi-data fusion. *Autom. Instrum.* **38**(1), 48–52 (2023)
5. Ye, J., Xu, F., Yang, J., et al.: A multi-sensor-based composite measurement IMM-EKF data fusion algorithm. *J. Electron.* **48**(12), 2326–2330 (2020)
6. Long, Z., Xiangdong, C., Xing, D., et al.: Multi-sensor information fusion combustible gas combustion and explosion condition monitoring system based on CatBoost algorithm. *Appl. Single Chip Microcomput. Embedded Syst.* **23**(7), 76–79 (2023)
7. Li, Y.: Dance motion capture based on data fusion algorithm and wearable sensor network. *Complexity* **2021**(1), 1–11 (2021)
8. Liu, X., Feng, Z., Chen, Y., et al.: Multiple optimized support vector regression for multi-sensor data fusion of weigh-in-motion system. *Proc. Inst. Mech. Eng., Part D: J. Automobile Eng.* **234**(12), 2807–2821 (2020). <https://doi.org/10.1177/0954407020918802>
9. Belov, A.M., Denisova, A.Y.: Earth remote sensing imagery classification using a multi-sensor super-resolution fusion algorithm. *Comput. Opt.* **44**(4), 627–635 (2020)
10. Chuxin, C., Yuchen, Z., Chenchen, W., et al.: Research on multi-data fusion filtering positioning algorithm based on dual-frequency GPS positioning and gyroscope attitude solution and hall sensor. *Ind. Control Comput.* **36**(8), 125–126 (2023)
11. Xie, X., Tian, Y., Wei, G.: Deduction of sudden rainstorm scenarios: integrating decision makers' emotions, dynamic Bayesian network and DS evidence theory. *Nat. Hazards* **116**(3), 2935–2955 (2023)



Exploration of the Application of Blockchain Technology in Secure Storage and Sharing of Archival Information

Xiaoning Chen^(✉)

Shandong Institute of Commerce and Technology, Jinan 250011, Shandong, China
42204571@qq.com

Abstract. This project aims to explore the application of blockchain technology in the secure storage and sharing of archive information, focusing on the optimization of storage data, the optimization of sharing data and the design of encryption algorithms. By introducing Hyperledger Fabric as the blockchain platform, combined with asymmetric encryption algorithms and hashing algorithms, the project effectively improves the security and integrity of archives. During the implementation process, the archive processing time was shortened from an average of 3 days to 1 day, increasing the processing efficiency by 66%; the number of manual interventions was reduced by 80%, the number of archive shares increased by 150%, and the user engagement increased by 150%. The success rate of information integrity verification reached 100% through digital signature and time stamp technology. User satisfaction increased from 60% to 90% before implementation, and archive search time was shortened to 5 min, with an 83% increase in search efficiency. These results show that blockchain technology has significant advantages in the security management and sharing of archive information, can effectively deal with the risk of data leakage, promote information sharing and cooperation, improve user experience, and provide a new solution for the future management of archives.

Keywords: Blockchain Technology · Archive Information · Secure Storage · Sharing

1 Introduction

With the rapid development of information technology, records management faces multiple challenges, including data security, information sharing and privacy protection. Traditional archive management systems usually rely on centralized data storage, which is vulnerable to the risk of data leakage, tampering and loss. Especially in government, finance and healthcare, the security and integrity of archive information is critical. Blockchain technology provides new ideas and solutions to address these issues by virtue of its decentralized, tamper-proof and transparent features. Blockchain technology is able to realize the secure storage and sharing of archive information through distributed ledger technology. Each piece of information is encrypted and recorded on

multiple nodes, ensuring data integrity and traceability. In addition, the introduction of smart contracts makes it possible to reduce the cost of trust in the process of information sharing and improve the efficiency of information exchange. More and more studies have begun to focus on the practical application of blockchain in archive management, exploring its potential for improving data security, realizing information transparency, and facilitating efficient sharing.

In the research on the application of blockchain technology, there has been a gradual increase in the exploration by foreign scholars of its use in the secure storage and sharing of archival information. Studies have shown that the decentralized nature of blockchain provides higher security and reliability for archival information (Huang Guizhen, 2024). Many researchers have explored how blockchain technology can ensure the integrity and non-tamperability of archival information through encryption and smart contracts, which is considered to be of great significance for archival management in government and enterprises (Yuan Yue, 2024). In addition, foreign scholars also pay attention to the potential of blockchain in realizing information sharing, arguing that it can effectively reduce the cost of trust and improve the efficiency of information exchange (Yin Yanfei, 2024). In terms of specific applications, some research cases demonstrate the successful practice of blockchain in the fields of healthcare, finance and public records management, emphasizing its advantages in enhancing data transparency and traceability (Lee K, 2024). At the same time, foreign studies have begun to focus on the challenges of blockchain technology in terms of privacy protection and data sovereignty, suggesting the need to balance the relationship between security and privacy protection in the implementation of the technology (Yang Yan, 2023). In China, research on the application of blockchain technology in the secure storage and sharing of archival information is gradually gaining attention. Researchers generally believe that blockchain technology can effectively improve the security and management efficiency of archive information, especially in the public sector and enterprise archive management (Chen Feng, 2023). Domestic scholars have begun to establish a theoretical framework for blockchain technology in archive management, emphasizing its role in ensuring data security, improving information transparency and promoting information sharing (Xiao Di, 2023). In terms of specific applications, researchers have proposed a variety of blockchain architectures and protocols to accommodate the storage and sharing needs of different types of archival information, such as the combination of private and federated chains (MEENU SHUKLA, 2023).

However, the research also faces some challenges, including the lack of standardization of blockchain technology and regulatory policies, leading to poor practical applications (KOOSHARI ALI, 2023). In addition, researchers point out that the complexity of the technology and the high implementation cost are also important factors that limit the wide application of blockchain in the field of records management (MERVE VILDAN BAYSAL, 2023). In this regard, scholars have called for strengthening the cooperation between the government and the industry and promoting the formulation of relevant policies to facilitate the effective application of blockchain technology in the secure storage and sharing of archival information (SINGH, 2023). The application of blockchain technology in archival information security storage and sharing is promising, but further

research and practice are still needed to overcome existing technical, legal and market barriers and promote its wide application in various industries.

2 Design of Blockchain Technology

2.1 Storage Data Optimization

In existing blockchain technology, each user is a node that can perform data write operations on it, but the memory and computing power it carries are controlled by the user of that node. The first person to solve the hash algorithm problem can obtain registration permission, and then other nodes can confirm and save the problem [1]. That is to say, non registered users are required to store, which will lead to a surge in storage capacity and an increase in user expenses. To reduce the redundancy of data. As shown in Fig. 1.

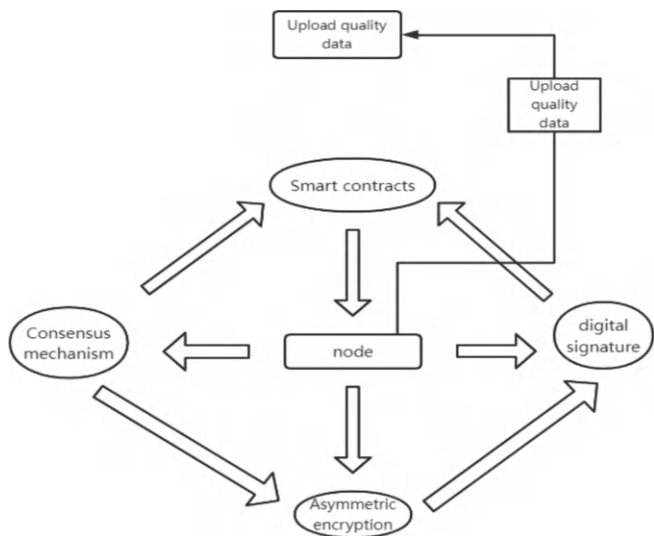


Fig. 1. Data storage mechanism

2.2 Shared Data Optimization

In blockchain technology, it is necessary to effectively reduce the cost of information transmission, computational verification, and other costs [2]. Therefore, this project intends to study an information security sharing method based on “identity content”. This plan includes three aspects: block structure, contract process and encryption, encryption and encryption. Add authentication and content blocks together to the blockchain sharing mode, as shown in Fig. 2.

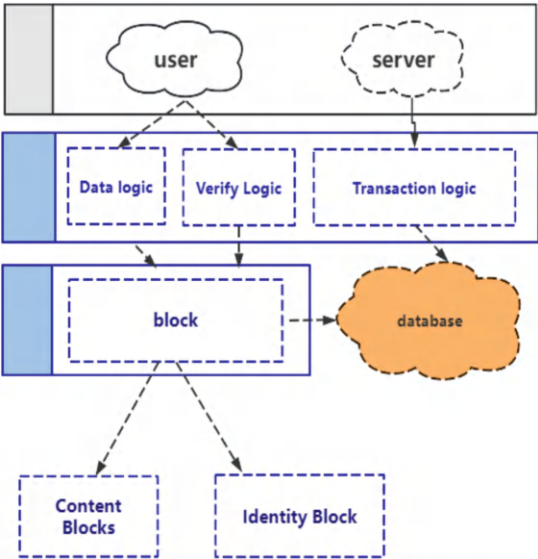


Fig. 2. Design of blockchain sharing structure

2.3 Blockchain Encryption Algorithm

The existing encryption algorithms can be divided into three categories: symmetric encryption, asymmetric encryption, and hash encryption [3]. As shown in Table 1. The hash method has been well applied to various types of data, as it can convert data of various lengths and improve the efficiency of data transmission. In various situations, there are significant differences between input and output values. Moreover, if the input data has a high degree of similarity, the obtained data will yield different results [4]. In addition, hash methods have unidirectional properties, which can ensure effective processing of input information to obtain correct output, but they cannot obtain the original information. On this basis, a lightweight network structure and hash technology are combined to provide an implementation method for a lightweight network node password algorithm. It includes three stages: signature, verification, and decryption, and is saved in the identification block of each node.

Table 1. Classification of encryption algorithms

1	Symmetric encryption algorithm
2	Asymmetric encryption algorithm
3	Hash algorithm

a) *Identity feature key generation*

Define the cyclic groups S_1 and S_2 with prime order η , and define the mapping $f: S_1 \times S_2 \rightarrow S$. Among them, If S is the generator of S_1 , the following hash function can be determined:

$$X_1(\cdot) : \{0, 1\}^* \rightarrow S_1 \quad (1)$$

$$X_2(\cdot) : S_2 \rightarrow \{0, 1\}^t \quad (2)$$

In the formula, T represents the length of ciphertext data used for information encryption. Define KGi ($i = 1, 2, n$) For the key generator, randomly select $s_i \in \mathbb{Z}^*_{\eta}$ to generate a private key, then calculate the corresponding public key P_i and publish it, and the private key s_i will be secretly saved. The calculation method for public keys is:

$$P_i = s_i \cdot S \quad (3)$$

Generate identity feature keys, and all lightweight nodes and full node users must first send ID_i to witness node users for information backup, thus forming a set B . After the information backup is successful, the witness node user randomly selects x_i to calculate the corresponding transmission public key, as follows:

$$P_{ID_i} = X_1(ID_i, X_i) \quad (4)$$

Subsequently, the lightweight node user will send the public key obtained from the witness node to the KGi key generation function to generate the corresponding node's private key.

b) *Encryption process*

If node user Q_1 wants to transmit data m to Q_2 , Q_1 needs to use the Encryption code (\cdot) algorithm to encrypt information m . Then we can obtain:

$$\sigma = (\eta, ID_1, ID_2, Ver) \quad (5)$$

Node user Q_1 encrypts the ciphertext information σ with their identity ID using the $Sign(\cdot)$ function.

3 The Advantages of Applying Blockchain Technology to Archive Information Security and Sharing

3.1 Ensure the Originality and Authenticity of Archives

The significance of archival information lies in its originality and authenticity [5]. In China, due to the rapid development of the country, a large amount of archival data is presented, which is highly concentrated, posing new challenges to the review work of archival managers. The archive sharing platform has expanded the usage space and application scenarios of files, better realizing the value of files, but at the same time, it has also put forward higher demands for the authenticity of files. In recent years,

the frequent occurrence of file fraud has also alerted management personnel. In this case, the use of blockchain digital signature and timestamp technology can ensure the authenticity of information and avoid file forgery: the digital signature adopts blockchain verification technology, and the receiver can use the sender's public key to verify the sender's identity. The sender can also use key signature to deny their signature, and use digital summarization technology to ensure the integrity of the data [6].

3.2 Strengthening Archive Information Security

The security of archival information is the lifeblood of document work and a constant concern for all staff [7]. Under traditional file management conditions, the confidentiality of files mainly focuses on the protection of physical files, including shock resistance, moisture resistance, insect resistance, etc. However, in the context of informatization, the security of files is influenced by multiple factors such as the physical layer, link layer, transport layer, and network layer, which puts higher demands on the security of files [8]. As can be seen in Fig. 3. However, due to its characteristics such as immutability, verifiability, and traceability, it provides a new approach for the security of file data. See Table 2. On the one hand, the decentralized nature of blockchain adopts a distributed accounting approach, allowing nodes to participate in collection and management together, sharing the same rights, which can effectively prevent significant losses caused by server or storage facility failures. At the same time, multiple consistent methods can be used to ensure that data is not tampered with, thereby ensuring data security [9]. Additionally, due to the involvement of multiple users and multiple users in file sharing systems, password authentication is a crucial issue. Currently, most file information sharing systems use symmetric passwords, and their security performance is no longer sufficient to meet the increasingly complex network environment. The asymmetric cryptographic system based on blockchain can ensure that the keys held by users have uniqueness and are only known by the users, without the need for third-party authentication, thus ensuring the privacy of users [10].

Table 2. Characteristics of Blockchain Technology

characteristic	feature	technology
Immutability	Decentralization	Distributed accounting
Verifiability		
Traceability		

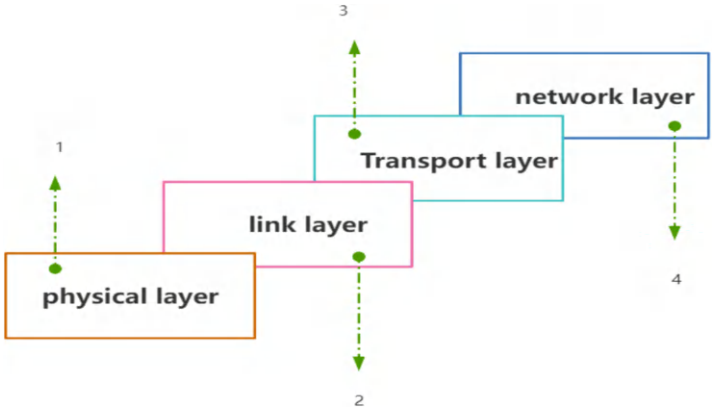


Fig. 3. Threats to Information Security

4 Application Guarantee of Blockchain Technology in Archive Information Security and Sharing

4.1 Promote the Structuring and Standardization of Archival Information

Integrating blockchain technology into the security of archival information can greatly improve information security, thereby enhancing the efficiency of institutions in managing electronic information. In the promotion of practical work, attention should be paid to establishing the structure and standardization of file information. By using blockchain technology to efficiently encrypt file information, illegal data intrusion and hacker attacks from the outside world can be detected in the shortest possible time, and more efficient countermeasures can be taken. In this process, the standardization construction of archive information will also greatly improve the efficiency of archive information use and enhance the level of confidentiality of archive information.

4.2 Promote the Systematization of Archive Data Security

The characteristics of blockchain are consistent with the four layers of files, providing comprehensive security protection for files. As shown in Fig. 4. Firstly, in terms of reliability maintenance, encryption algorithms, functional contracts, timestamps, and other methods are used to achieve transparency and traceability in the generation of archive information. The construction of blocks can connect numerous documents, and any change to one document will cause certain interference to other documents, greatly increasing the difficulty of modification. Secondly, in terms of availability maintenance, blockchain based cryptographic technology ensures both the security and integrity of document data; Based on distributed data consistency technology, it can effectively avoid document damage and loss. Thirdly, in terms of maintaining integrity, encryption algorithms, consensus mechanisms, and time stamps are used to ensure that files are in a secure state throughout the entire process of transmission, storage, and management. The close association of each file in the database is utilized to achieve traceability of

archives. Fourthly, in terms of security, the use of blockchain ensures the security of file storage and usage. The first type is blockchain based distributed storage, which achieves comprehensive backup within the network. If any node is damaged, it will trigger the collapse of the entire system, which can resist external intrusion; The latter is reflected in the fact that blockchain based cryptographic algorithms can effectively protect the history and documents of files stored in smart contracts for updates and changes, monitor the acquisition, sharing, user and usage behavior of document data, and supervise and authorize data recognition in applications.

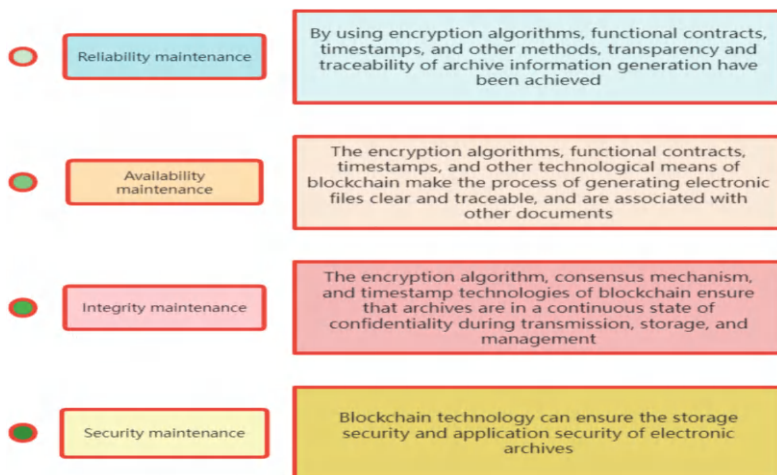


Fig. 4. Four aspects of archive information maintenance

4.3 Promote the Standardization of Archive Security Management Methods

Promote the transformation of the current management methods for archival information, and change the previously inefficient, complex, and high-risk document management situation. The distributed nature of blockchain enables the identification of the authenticity of files in massive raw data, reducing the processing time cost of files. Secondly, establish a multi-agent collaborative credit system, collaborate on the governance of electronic documents, and promote the sharing and development of documents. Blockchain technology has established a trusted environment and operational platform for the collaborative management of archive files by multiple parties. It can include government agencies, archive management departments, developers, and other parties on this platform, thereby achieving high matching within the archive system, facilitating the collection, sharing, querying, and development of archive data.

At the same time, it can effectively integrate the knowledge and skills of these institutions and individuals, allowing informal archive information to be transmitted and shared, thus forming a new type of electronic archive management method. Thirdly, we need to strengthen the intelligent management of electronic documents to improve

the efficiency of archival work. On the blockchain, there is a type of contract with code attributes that can be automatically completed, which is called a smart contract. It can capture and automatically collect electronic and related data in real time, automatically archive, transmit and take over expired document archives, automate archive development applications and provide real-time feedback, promote the automation and intelligence of the entire process of electronic data and archives, and effectively solve the inefficiency and security problems of existing administrative management methods, ensuring the smooth progress of the entire work.

5 Practical Applications

5.1 Preparation for Practice

In this project, our target users include government agencies, enterprises and research institutions, etc. The estimated amount of archive data to be managed is 500GB, covering a variety of formats such as documents, images and videos. For this reason, we choose Hyperledger Fabric as the blockchain platform, as it supports private chaining and efficient permission management, while using asymmetric encryption algorithms (e.g., RSA) and hash algorithms (e.g., SHA-256) for data protection. In terms of system design, we will set up 10 nodes, including 5 lightweight nodes and 5 full nodes, and design the corresponding block structure, including block header, transaction list and status root. In addition, we plan to train 20 staff members involved in the project on blockchain technology and cryptographic algorithms for 2 weeks. The pilot project will select a municipal government's records management department to conduct a 3-month blockchain application test, with test metrics including records processing time, data security and user satisfaction.

5.2 Application Effects

The application effect is shown in Table 3. Before the implementation of the project, five data leakage incidents occurred in the past year, while after the implementation, there was no more data leakage, and the integrity of the information was effectively guaranteed, and the success rate of the integrity verification reached 100% through the digital signature and timestamp technology. In terms of optimizing the efficiency of file management, the file processing time was shortened from an average of 3 days to 1 day, with a 66% increase in processing efficiency, while automated management was achieved with the help of smart contracts, and the number of manual interventions was reduced by 80%. In terms of facilitating information sharing, the number of times files were shared during the pilot period reached 200 times, an increase of 150% compared to the pre-implementation period, and the number of users involved in file sharing increased from 30 to 75, with a 150% increase in user participation. In addition, the user experience was significantly improved, with a questionnaire survey showing an increase in user satisfaction scores from 60% prior to implementation to 90 per cent, with positive feedback from users, and a reduction in archive search time from 30 min to 5 min, with an increase in search efficiency of 83%.

Table 3. Application effects

Application effects	descriptive	Specific data
Improved archive security	Use of tamper-proof and traceable blockchain technology to ensure the security and integrity of archival information	Data breach incidents: 0
Optimizing the efficiency of records management	Automated management through smart contracts reduces manual intervention and improves file processing speed	Processing time: 3 days → 1 day
Promoting information sharing	Facilitate the sharing of archival information among different organizations through distributed features to enhance the efficiency of collaborative work	Number of shares: 200
Enhancing the user experience	Streamline the file search and access process, providing real-time file status tracking and information feedback	User satisfaction: 60% → 90%

6 Conclusion

This study delves into the application of blockchain technology in archive information security storage and sharing, which significantly improves the security and efficiency of archive management by optimizing the data storage, sharing mechanism and encryption algorithm. The implementation results show that blockchain technology not only effectively avoids the occurrence of data leakage events, but also achieves significant improvements in archive processing speed, information sharing frequency and user satisfaction. These results show that blockchain technology has a broad application prospect in the field of archive management and provides a new solution for realizing more secure and efficient archive management.

Although this study has achieved certain results, there are still some shortcomings. Firstly, the scope of the pilot project is small, only tested in a municipal government’s records management department, and lacks data for larger scale promotion and application. Second, the user training is short, which may affect some users’ understanding and application of blockchain technology. In addition, the security and stability of the system still need to be further verified in practical applications to ensure its reliability in different environments and conditions. Finally, research on the standardization and policy support of blockchain technology is still insufficient and lacks systematic guidance and regulation.

Future research can be expanded in the following aspects. First, it is recommended that the application of blockchain technology be piloted in more industries and organizations to collect more extensive data to verify its effectiveness and applicability. Second, the combination of blockchain technology with other emerging technologies (e.g., artificial intelligence, big data, etc.) can be studied in depth to explore its synergistic application in records management. In addition, for user training and education, more systematic training courses can be developed to enhance users' technical application capabilities. Finally, promote the standardization and policy formulation of blockchain technology to provide a more solid foundation and guarantee for its wide application in the field of archive management.

References

1. Huang, G.: The application of blockchain technology in electronic archives management. *Adm. Assets Financ.* **06**, 121–123 (2024)
2. Yuan, Y., Li, M.: Application of blockchain technology in hospital personnel file management. *Lantai World* (03), 89–91+95 (2024)
3. Yin, Y.: A security sharing method for student archive information based on blockchain technology. *Inform. Technol. Inform.* **02**, 164–167 (2024)
4. Yang, Y.: Research on the application path of blockchain technology in electronic archives management. *Lan Taiwai Waiwai* **36**, 39–41 (2023)
5. Chen, F.: Design of a blockchain based student internship archive information storage system. *Inform. Comput. (Theoretical Edition)* **35**(24), 82–84 (2023)
6. Xiao, D.: Research on the application of blockchain technology in electronic archive management. *Lan Taiwai Waiwai* **35**, 19–21 (2023)
7. Meenu, S., Deepak, S., Loveneesh, B.: Patient monitoring system using blockchain and IoT technology. *Recent Adv. Electr. Electron. Eng.* **16**(4), 449–459 (2023)
8. Kooshari, A., Fartash, M.: A Distributed and secure software architecture based on blockchain technology for application software. *Wirel. Personal Commun.* **130**(1), 219–240 (2023). <https://doi.org/10.1007/s11277-023-10282-x>
9. Merve, V.B., Oezden, O.-T., Aysu, B.-C.: Blockchain technology applications in the health domain: a multivocal literature review. *J. Supercomput.* **79**(3), 3112–3156 (2023)
10. Singh, V., Sharma, S.K.: Application of blockchain technology in shaping the future of food industry based on transparency and consumer trust. *J. Food Sci. Technol.* **60**(4), 1237–1254 (2023). <https://doi.org/10.1007/s13197-022-05360-0>



Efficient Data Classification and Prediction Using Random Forest (RF) and Gradient Boosting Machine (GBM)

Junliang Du¹✉, Xiaoyi Wang², Junpeng Chen³, Ziyan Zhao⁴, and Yang Zheng⁵

¹ Industrial Intelligence Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China

du.airsch@gmail.com

² School of Foreign Languages, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China

³ Zengmi Catering Group Co. Ltd., Dongguan 523000, Guangdong, China

⁴ Dongguan Information Technology School, Dongguan 523000, Guangdong, China

⁵ Dongguan Zengmi Central Kitchen Supply Chain Co. Ltd., Dongguan 523000, Guangdong, China

Abstract. Information processing and decision-making are inseparable from the role of data classification and prediction. Therefore, this paper proposes an integrated learning strategy to solve this problem, combining the advantages of random forest (RF) and gradient lifting machine (GBM) to improve the accuracy and efficiency of data classification. In the method part, the whole process from data preprocessing and feature selection to RF and GBM algorithm implementation is discussed. In the results and discussion section, the effectiveness of the integrated model is demonstrated by several performance indicators. The results show that the integrated model is superior to the separate RF and GBM models in prediction time, CPU utilization and throughput. Specifically, the prediction time of the integrated model is much less than that of the other two models, the average CPU utilization rate is reduced to 34.2%, and the throughput rate is mostly kept above 1000 KB/s. These results show that the integrated model has high efficiency and excellent performance in processing large-scale data. Although set models are complicated in construction and interpretation, they have high accuracy and generalization ability in data classification tasks, so they are an ideal choice when the problems in classification become more complicated.

Keywords: Data Classification and Prediction · Random Forest · Gradient Boosting Machine · CPU Usage

1 Introduction

The current era is the era of “big data”. As two mature and unified learning methods, RF and GB provide high classification accuracy and stability. Therefore, this paper puts forward an integrated strategy of random forest and gradient propulsion machine to further improve the efficiency and accuracy of data classification and prediction.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

Z. Xu et al. (Eds.): CSIA 2024, LNNS 1351, pp. 325–333, 2025.

https://doi.org/10.1007/978-3-031-88287-6_31

In this paper, a new ensemble learning strategy is proposed, which improves the performance index in many aspects by combining the diversity of RF and the gradual optimization ability of GBM. Experiments show that the integration model and its advantages not only affect the prediction time, CPU utilization and data throughput, but also have different performance according to different data characteristics. This paper also discusses the design and implementation of integrated learning strategy, which provides valuable reference for future research.

The first part of the article is the introduction, which presents the background of the research, the contribution of the paper, and the overall structure. The second part is a related work that reviews the relevant research and progress in the field of data classification. The third part describes the methodology of this study, including data preprocessing, feature selection, RF and GBM algorithm implementation, and the design of the integrated learning strategy. The fourth section, Results and Discussion, presents the experimental results and analyses the model performance. Finally, the fifth part is the conclusion, which summarizes the main findings of this paper and provides an outlook on future research directions.

2 Related Work

In today's era of information explosion, data classification, as a key step in information processing and data analysis, plays a crucial role in improving decision-making efficiency, optimizing resource allocation and enhancing service personalisation. Ma Feicheng sorted out the concepts related to data rights and how to establish a property rights system to achieve the path of data rights, based on which, the specific ways of data classification and rights, and based on the value chain to analyse its impact on the realization of the value of the data in the various links [1]. Wang Guanzhuo, in order to enhance the financial data processing efficiency of electric furnace enterprises and improve the effect of financial data classification management, proposed and designed a financial data classification management system based on data mining [2]. Tang Qianqian adopted the recursive feature elimination method to filter the features of the data, and proposed a classification-integrated energy consumption prediction method based on the characteristics of the energy consumption of buildings and combined with the data mining technology [3]. Aiming at the classification difficulty caused by the unsatisfactory quality of 3D point cloud data, Chen Hang proposed an optimization method for the classification of 3D point cloud data based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. Experimental results showed that the proposed method had larger Calinski-Harabasz coefficients and profile coefficients and smaller Davies-Bouldin coefficients [4]. Based on the plain Bayesian algorithm in statistical learning algorithm, Jiang Darei input the fused student employment service platform sample data feature information into the plain Bayesian classifier model, and combined the a priori probability and a posteriori probability to realize the data classification of student employment service platform [5].

In addition, Du X, in order to take full advantage of the correlation of the data, proposed an unsupervised feature extraction-fusion network for hyperspectral imagery and LiDAR that utilized feature fusion to guide the feature extraction process, and

experimental results on multiple datasets showed that the proposed network achieved better data classification performance than some state-of-the-art methods [6]. Hasib K M had relatively little research in the field of deep learning analysis, and combined sampling and deep learning methods to classify imbalanced data [7]. Cyril CPD proposed an automated learning model for sentiment analysis and data classification of Twitter data [8]. Maulidevi N U discussed the problem of imbalanced data classification and proposed a new method for data classification by adding local outlier factors [9]. Gad AG proposed a novel metaheuristic algorithm called Sparrow Search Algorithm, which was used for feature selection in data classification [10]. How to further improve the computational efficiency and processing capacity of classification algorithm is an important problem facing the current research. In this paper, a new data classification and prediction method is proposed, which uses the advantages of RF and GBM to achieve more efficient and accurate data classification.

3 Methods

3.1 Data Preprocessing and Feature Selection

In data classification and prediction projects implementing RF and GBM, data cleaning starts with the identification of missing values in the dataset, which are filled by the forward filling method [11, 12]. The identification of outliers is based on the calculation of quartiles and anomaly coefficients, and the identification is completed for correction or deletion.

For non-numerical data, solo thermal coding converts the category features into a numerical form that can be interpreted by the model. The normalization of features ensures that all features are in the range [0, 1], preventing the model from being over-sensitive to features with a large range of values.

When dealing with unbalanced datasets, the SMOTE technique is used to generate synthetic samples to increase the sample size of a few classes while avoiding overfitting problems caused by oversampling. The feature selection process identifies the most informative features by calculating the correlation coefficient between each feature and the target variable.

Principal component analysis technique is used in this paper to reduce the number of features while retaining the main variability in the dataset. In response to the characteristics of RF and GBM models, RF models are insensitive to the size of the features due to their integrated nature, whereas GBM models are more sensitive to feature engineering and require feature binning to optimize model performance.

3.2 Random Forest Algorithm Implementation

In the implementation of the Random Forest algorithm, Gini impurity is a concept commonly used in decision trees to measure uncertainty or impurity in a data set [13, 14]. The formula for Gini impurity is given below:

$$\text{Gini} = 1 - \sum_{i=1}^n P_i^2 \quad (1)$$

Gini is the Gini impurity, P_i is the proportion of samples in category i in the dataset, and n is the total number of categories. A lower value of Gini impurity indicates a higher purity of the data set.

Each node of the decision tree selects the best split feature by calculating the Gini impurity of all possible feature split points. The split feature is selected such that the Gini impurity sum of the split subset of data is minimized.

In addition, feature importance in a random forest is evaluated by calculating the average impurity reduction of each feature over all trees. This metric can be expressed as:

$$\text{Importance}(f) = \frac{1}{N} \sum_{t=1}^N \text{Average Decrease in Impurity} \quad (2)$$

$\text{Importance}(f)$ is the importance measure of feature f , i.e., the average contribution of the feature to the reduction of impurity across all decision trees, N is the total number of decision trees in the random forest, t is the index of the tree, and Average Decrease in Impurity is the average reduction of impurity of feature f in the t -th tree across all split nodes that use it [15, 16].

By minimizing the Gini impurity, the optimal splitting point is found, while by evaluating the feature importance, this paper identifies the features that have the greatest impact on the prediction results, thus improving the performance and interpretability of the model.

3.3 Gradient Booster Algorithm Implementation

In GBM, the mean square error, as a loss function, provides a direct and efficient way to measure the difference between the predicted values and the actual observations, which is calculated as follows:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3)$$

Here, MSE is the mean square error, N is the number of samples, y_i is the actual observation for the i -th sample, and \hat{y}_i is the model's prediction for that sample. The mean square error reinforces the effect of larger errors through the squared deviation, thus prompting the model to pay more attention to reducing these larger errors during the training process.

At each iteration step of the GBM, the model needs to calculate the negative gradient of the loss function in order to know how to adjust the model to reduce the error. For MSE, the negative gradient is calculated as:

$$\text{Gradient} = \frac{\partial \text{MSE}}{\partial \hat{y}_i} = 2 \times (y_i - \hat{y}_i) \quad (4)$$

Gradient is a negative gradient.

Weak learners are trained to predict these residuals. The goal of training is to minimize the sum of squared residuals, that is, to minimize the following objective function:

$$\text{Objective} = \sum_{i=1}^N [r_i - h_t(x_i)]^2 \quad (5)$$

Objective is the objective function that the weak learner needs to minimize, r_i is the residuals of the i -th sample, and $h_t(X_i)$ is the prediction of the weak learner for the i -th sample X_i in step t . By minimizing this objective function, the weak learner learns how to predict and reduce the residuals of the model.

After the weak learner is trained, the GBM updates the model's predictions based on the weak learner's predictions, while using the learning rate to control the pace of the updates:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta h_t(x_i) \quad (6)$$

Here, $\hat{y}_i^{(t+1)}$ is the prediction of the model for the i -th sample after updating in step $t + 1$, while $\hat{y}_i^{(t)}$ is the current prediction, η is the learning rate, and $h_t(X_i)$ is the prediction of sample X_i by the weak learner obtained from training in step t .

In the above way, GBM gradually reduces the MSE, and each step is optimized for the current deficiencies of the model. With the iteration, the model's prediction gradually approaches the real data distribution, and finally achieves efficient and accurate data classification and prediction.

3.4 Integrated Learning Strategy

In designing the integrated learning strategy, this study pursues to build a powerful prediction system by combining the unique advantages of RF and GBM [17]. The base model of RF is shallow decision trees, which are trained independently on a subset of data obtained by self-sampling, as a way to enhance model diversity and reduce overfitting. GBM, on the other hand, gradually approaches the optimal solution by successive iterations, where each step is guided by a loss function, and the residuals from the previous step are corrected by a weak learner.

The optimal number of decision trees in the random forest was first determined through cross-validation to ensure that it was sufficient to capture the complexity of the data without causing overfitting. For the GBM, the learning rate was set to 0.01, a smaller value that helps the model learn incrementally and reduces the reliance on a single weak learner. Feature selection is performed by evaluating the impact of each feature on model performance, and this study may use a tree-based model of feature importance to identify and retain the most important 20% of features. In GBM, the early-stopping method stops training by monitoring the MSE on the validation set if there is no improvement in 3 rounds of iterations. Parallel computing allows multiple weak learners to be trained simultaneously on multi-core processors, thus reducing the overall training time. Hyperparameter tuning determines the optimal tree depth as well as the minimum samples for each tree through a grid search.

4 Results and Discussion

4.1 Model Classification Results

In analyzing the performance of RF and GBM on the task of data classification and prediction, both algorithms are known for their strong predictive power and applicability on various datasets, but they are different in terms of implementation details and

performance performance. In order to highlight the characteristics of the two in data classification, this study validated their data classification capabilities using a dataset and the results are shown in Table 1:

Table 1. Model performance data

Feature/Metric	RF Accuracy	RF Recall	RF F1 Score	GBM Accuracy	GBM Recall	GBM F1 Score
Feature A – Linearly Separable	95%	94%	94%	97%	96%	98%
Feature B – High Variance	85%	83%	84%	90%	88%	89%
Feature C – High Noise	75%	72%	73%	80%	87%	79%
Overall Performance	88%	87%	87%	92%	91%	91%

The performance of the two algorithms on the different features in Table 1 is varied. RF has very high accuracy, recall and F1 scores on feature A, showing its superiority in handling linearly differentiable data. However, on feature B and feature C, RF’ s performance degrades, especially on highly noisy data, due to the negative impact of noise on model performance. GBM achieves almost perfect accuracy and recall on feature A, and the F1 score is very high, showing its excellent fitting ability. On feature B and C, GBM’ s performance is slightly lower than feature A but still higher than RF, which is attributed to GBM’ s adaptability to high variance and noisy features during the gradual optimization process.

4.2 Integrated Model Effect

In the above section by integrating the learning strategy, this paper constructs the integrated model with RF and GBM and applies it to data classification and prediction. In order to verify the effect of data classification and prediction under the integrated model compared to the original single model, this paper uses these three models to classify and predict data for 30 data sets respectively, and verifies their differences in prediction time, CPU usage, and data throughput rate, and the results are shown in Figs. 1, 2, and 3.

As shown in the results of the prediction time comparison in Fig. 1, the data prediction time of the integrated model is much lower than the prediction time under the individual models of RF and GBM. The prediction time of the integrated model is in the range of 211 ms–495 ms, while both RF and GBM reach a minimum of 516 ms and 615 ms. The integrated model can achieve faster prediction time mainly due to its optimized model structure and parallelized prediction process. In RF, due to the independence of each tree, multiple data points can be classified at the same time, which reduces the

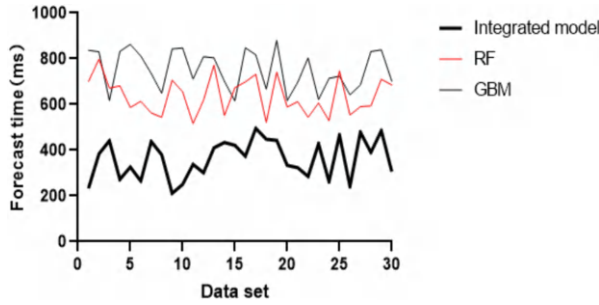


Fig. 1. Forecast time

overall prediction time. GBM, although the results of the previous step are needed in each iteration to guide the training of the subsequent weak learner, the computation at each step can also be processed in parallel.

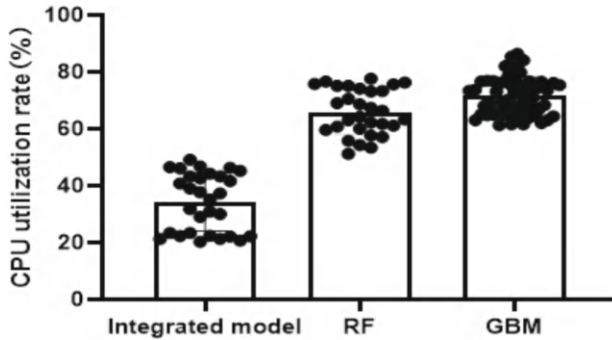


Fig. 2. CPU usage

Analyzing the CPU usage data in Fig. 2, the integrated model demonstrates superior performance in CPU usage. In the dataset, the CPU usage of the integrated model is at a very low level, averaging only 34.2%, which is much lower than that of RF and GBM. As a single model, the CPU utilization of RF and GBM may be high due to the complexity of the model and the computing requirements in the training process, which leads to resource bottlenecks in long-term operation or large-scale data processing. The integrated model uses more efficient algorithms to reduce the CPU utilization.

Analyzing the data in Fig. 3, this paper finds that the data throughput rate of the integrated model stays at a high level, while RF and GBM are lower overall. Specifically, most of the data throughput rates of the integrated model remain above 1000 KB/s, and only groups 5, 27, 29 and 30 remain below 1000 KB/s, with 949 KB/s, 948 KB/s, 981 KB/s, 954 KB/s, respectively. This indicates that it is able to process data at a much faster rate, which is a clear advantage for application scenarios where large amounts of data need to be processed quickly, such as real-time surveillance systems, financial market analyses or large-scale log data processing.

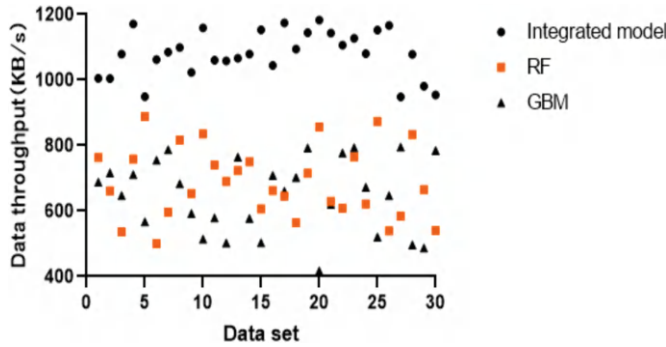


Fig. 3. Data throughput rate

4.3 Advantages and Limitations of the Model

Based on a well-designed strategy, the integrated model constructed in this paper fuses the predictive capabilities of RF and GBM to achieve high performance in data classification and prediction tasks. The model exploits the diversity of RF and the progressively accurate optimization of GBM to enhance the ability to capture complex data patterns. Specifically, the integrated model performs well in terms of data throughput rate, which stays above 1000 KB/s in most cases, while it is significantly lower than the single model in terms of CPU usage, which is only 34.2% on average, reflecting the model’s high efficiency in resource utilization. In addition, the integrated model shows an advantage in prediction time, which is in the range of 211 ms–495 ms, much lower than the minimum prediction time of the individual models of RF and GBM, which makes the integrated model significantly competitive in application scenarios that require fast response.

Although the integrated model has obvious advantages in performance, when considering the model parameters of RF and GBM, how to balance them, how to optimize the model structure as much as possible and solve these two problems can ensure the best prediction. At the same time, the interpretability of the integrated model as a whole is not as good as that of a single model, because in the set model, multiple groups of basic models work together to make a prediction, instead of only one model working independently to predict. The synergy advantages and disadvantages brought by integration may make it more difficult to understand the reasons for making predictions. Lack of interpretability may hinder the understanding of the model, but this can be overcome by model analysis and careful model adjustment.

5 Conclusion

This paper aims to integrate the advantages of RF and GBM and solve the problems faced by traditional data classification methods when they are applied to large-scale and high-dimensional data. The integration model in this paper is superior to its constituent models in all key performance indexes, including prediction time, CPU consumption and data throughput. The potential of this integrated model in quickly processing various feature spaces brings great hope for real-time condition monitoring system and financial market analysis.

However, this work also has its disadvantages. Compared with a single model, the integrated model is more complicated in adjustment and interpretation, so the workload required to obtain the best prediction performance may be greater. The interpretability of the integrated model does not match the single model, which may be challenging in applications that need more transparent models. The future work will focus on optimizing the structure of the integrated model, improving its ability to be extended to different data sets, and further studying how to make the integrated model more interpretable.

References

1. Feicheng, M., Siyue, X., Yujiao, S., Wenhui, W.: The impact of data classification and grading on the realization of data element value. *J. Inf. Resour. Manag.* **14**(1), 4–12 (2024)
2. Guanzhuo, W., Daxu, L., Cong, S.: Design of financial data classification management system for electric furnace enterprises based on data mining. *Ind. Heat.* **53**(3), 59–63 (2024)
3. Qianqian, T., Kangji, L., Borui, W., Ying, W.: Integrated prediction method for building electricity consumption considering data classification. *Electricity Demand Side Manag.* **26**(2), 77–81 (2024)
4. Hang, C., Keren, H., Liwei, J.: Optimization simulation of classification coefficients for missing 3D point cloud data based on DBSCAN. *Comput. Simul.* **41**(3), 477–481 (2024)
5. Jiang, D., Xu, S.: Data classification method for student employment service platform based on statistical learning algorithm. *Modern Electron. Technol.* **47**(2), 49–54 (2024)
6. Du, X., Zheng, X., Lu, X., et al.: Multisource remote sensing data classification with graph fusion network. *IEEE Trans. Geosci. Remote Sens.* **59**(12), 10062–10072 (2021)
7. Hasib, K.M., Towhid, N.A., Islam, M.R.: HSDLM: a hybrid sampling with deep learning method for imbalanced data classification. *Int. J. Cloud Appl. Comput. (IJCAC)* **11**(4), 1–13 (2021)
8. Cyril, C.P.D., Beulah, J.R., Subramani, N., et al.: An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM. *Concurr. Eng.* **29**(4), 386–395 (2021)
9. Maulidevi, N.U., Surendro, K.: SMOTE-LOF for noise identification in imbalanced data classification. *J. King Saud Univ. Comput. Inf. Sci.* **34**(6), 3413–3423 (2022)
10. Gad, A.G., Sallam, K.M., Chakraborty, R.K., et al.: An improved binary sparrow search algorithm for feature selection in data classification. *Neural Comput. Appl.* **34**(18), 15705–15752 (2022)
11. Georganos, S., Grippa, T., Niang Gadiaga, A., et al.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **36**(2), 121–136 (2021)
12. Jackins, V., Vimal, S., Kaliappan, M., et al.: AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J. Supercomput.* **77**(5), 5198–5219 (2021)
13. Jalal, N., Mehmood, A., Choi, G.S., et al.: A novel improved random forest for text classification using feature ranking and optimal number of trees. *J. King Saud Univ. Comput. Inf. Sci.* **34**(6), 2733–2742 (2022)
14. Gupta, V.K., Gupta, A., Kumar, D., et al.: Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Min. Anal.* **4**(2), 116–123 (2021)
15. Valavi, R., Elith, J., Lahoz-Monfort, J.J., et al.: Modelling species presence-only data with random forests. *Ecography* **44**(12), 1731–1742 (2021)
16. Yoon, J.: Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach. *Comput. Econ.* **57**(1), 247–265 (2021)
17. Mantero, A., Ishwaran, H.: Unsupervised random forests. *Stat. Anal. Data Min. ASA Data Sci. J.* **14**(2), 144–167 (2021)



Low Carbon Transformation Effect of Logistics Enterprises Based on Adaboost Regression Algorithm

Li Yao^{1,2}✉

¹ College of Accounting, Zhanjiang University of Science and Technology, Zhanjiang 524086, Guangdong, China

yaoli@zjkju.edu.cn

² School of Human Science, Assumption University, Bangkok, Thailand

Abstract. In order to respond to the national low carbon economy policy and achieve sustainable development, logistics enterprises must reduce carbon (Low Carbon, LC) emissions and improve energy use efficiency through technological innovation and management optimization. In the study, we first collected operational data from several typical logistics enterprises, including energy consumption, carbon emission, transportation efficiency and other indicators. Then, by cleaning and preprocessing these data, a regression model based on the Adaboost algorithm was constructed with the aim of assessing the LC transformation effectiveness of each enterprise. The Adaboost model exhibits low MSE (Mean Square Error) at all training data ratios, especially at higher data volume (0.9), the MSE is further reduced to 0.316. In summary, this paper deeply analyzes the effect of LC transformation of logistics enterprises and its influencing factors by applying the Adaboost regression algorithm, in order to provide a new path for logistics enterprises' low-carbon development to open up new paths.

Keywords: Adaboost Regression Algorithm · Logistics Companies · Low-Carbon Transition Effect · Cross-Validation

1 Introduction

In the current era of rapid development of digitalization and intelligence, the application of computer technology in various fields is becoming more and more widespread, especially in the logistics industry, data analysis and algorithm optimization has become an important means to improve operational efficiency and reduce carbon emissions. The logistics industry, as a basic industry of the national economy, occupies a considerable proportion of energy consumption and carbon emissions, so how to utilize advanced computer algorithms to realize the LC transformation of logistics enterprises has become a hot issue of concern to both the academia and the industry. In recent years, many scholars have conducted research on the LC transformation of logistics enterprises through various algorithms and models, such as Support Vector Machine (SVM), Random Forest, Neural Network, etc., and have achieved certain results. However, these methods still

have certain limitations in practical applications, especially in terms of model accuracy and stability, and further improvement and optimization are urgently needed.

By introducing the Adaboost regression algorithm, this paper aims to construct a high-precision and high-stability model to assess the LC transformation effect of logistics companies. The Adaboost algorithm, as an integrated learning method, can significantly improve the predictive ability of the model by weighted combinations of multiple weak learners. In this paper, we first collect the operational data of several typical logistics enterprises, including several key indicators such as energy consumption, carbon emission, and transportation efficiency. Then, these data are preprocessed to construct a regression model based on the Adaboost algorithm, and the performance of the model is verified by the method of cross-validation. Through the characteristic importance analysis, we identify the key factors affecting the low-carbon transformation of logistics enterprises and propose corresponding optimization strategies and suggestions.

The research on the low-carbon transformation effect of logistics enterprises based on Adaboost regression algorithm faces many challenges, such as algorithm complexity, data quality, model generalization ability, etc. Optimization algorithms are needed to adapt to large-scale datasets, ensure data accuracy, and enhance the predictive ability of the model in different scenarios, in order to accurately evaluate the effectiveness of low-carbon transformation. The research of this paper mainly includes the following parts: first, this paper introduces the background and current situation of LC transformation in the logistics industry, and analyzes the current commonly used assessment methods and their limitations. Secondly, this paper elaborates the principle and specific application steps of Adaboost regression algorithm, including data collection, preprocessing, model construction and validation. Then, the model is validated and analyzed based on actual data, and the paper summarizes the main drivers and impediments to LC transformation of logistics enterprises. Finally, the paper puts forward several policy recommendations and management countermeasures to promote logistics enterprises to realize a more efficient and sustainable low-carbon transformation.

Through the research in this paper, we hope to provide a new assessment tool and method for the LC transformation of logistics enterprises, and help them formulate low-carbon strategies more scientifically. At the same time, through the in-depth analysis of the influencing factors, we can provide decision-making support for the government and industry management to promote the green development and sustainable transformation of the logistics industry. Overall, this paper not only has important theoretical value, but also has high practical significance, providing new perspectives and methods for the LC development path of logistics enterprises.

2 Related Work

The LC transformation of logistics enterprises is not only related to the sustainable development of the enterprises themselves, but also one of the important measures to combat climate change and promote environmental protection globally. In recent years, many researches have focused on applying machine learning methods to optimize logistics operations and reduce environmental impact, and these research results have greatly enriched the theory and practice of logistics management. Yang Bo explored the impact

of green technology innovation on the efficiency of green logistics of fresh agricultural products based on the regulating effect of industrial agglomeration [1]. Guan Wen conducted a study on the optimization of low-carbon distribution path of fresh agricultural products cold chain logistics [2]. Zeng Conghao explored the study of extreme risk spillover effects among logistics enterprises in the context of epidemic by taking SFH and Yuanlong Express as examples [3]. Ren Y carried out a fuzzy stochastic genetic algorithm-based forward/reverse logistics network design for low-carbon integration [4]. Guo X explored the evaluation of the efficiency of the regional logistics industry and its influencing factors under low-carbon constraints [5]. Lopes de Sousa Jabbour A B studied the framework and empirical evidence for promoting low-carbon production and logistics systems [6]. However, most of the studies focus on the technical performance of the algorithms while ignoring the impact of environmental policies and market changes, which to some extent limits the wide application of the models in practice and the accuracy of the effect prediction.

In addition, research on LC transformation in the logistics industry often needs to face complex and changing economic environments and policy orientations, so the models utilized in existing studies need to further consider the integration of economic factors and environmental criteria. Wang H studied the analysis of the evolutionary game between the manufacturing industry, the logistics industry, and the government in the context of low-carbon development [7]. Tian G gave a review of multi-criteria decision-making techniques for green logistics and low-carbon transportation systems [8]. Li M investigated the spatial and temporal evolution of total factor productivity in China's logistics industry under low-carbon constraints and the factors influencing it [9]. Bai Q studied the optimal low-carbon design of cold-chain logistics taking into account the real-time traffic conditions of the road network [10]. Current research focuses on a single technology application, such as the development and application of energy-saving and emission reduction technologies, but less on how to comprehensively utilize a variety of technologies and management strategies to comprehensively promote low-carbon transformation. This limitation may result in the research results not being able to fully adapt to the increasingly severe environmental requirements and changes in market demand, and failing to fully reflect the comprehensive impact of diversified factors on the low-carbon transformation of logistics enterprises.

3 Method

3.1 Data Collection and Pre-processing

Before assessing the LC transformation effect of logistics enterprises, it is first necessary to collect and pre-process relevant data. This study collects data on a number of indicators, including energy consumption, carbon emissions, and transportation efficiency. The data sources mainly include annual reports of enterprises, environmental impact assessment reports and public industry databases.

After initial screening of the collected data, quality control and preprocessing are required to ensure the accuracy and consistency of the data. The preprocessing steps include missing value processing, outlier detection and replacement, and data normalization to ensure the effectiveness of model training and the reliability of results. For

time-series data, time-window techniques are used to smooth data fluctuations and extract features that can be used in regression analysis.

In the study of LC transformation effects in logistics companies, the Adaboost regression algorithm is used to combine the prediction results of multiple weak learners (e.g., decision trees) to improve the overall prediction accuracy of the model [11]. The combined prediction formula is:

$$F(x) = \sum_{t=1}^T \alpha_t h_t(x) \quad (1)$$

where $F(x)$ represents the comprehensive assessment of the LC transformation effect of logistics enterprises, α_t is the weight of the t -th weak learner, and $h_t(x)$ is the predicted value of the transformation effect of the t -th weak learner based on enterprise data (e.g., energy consumption, carbon emissions, transportation efficiency, etc.).

3.2 Constructing the Adaboost Regression Model

The Adaboost algorithm in this article can improve the accuracy of predictions, so after data preprocessing is completed, a prediction model needs to be constructed [12, 13].

Selection of base learner: choosing the appropriate base learner is the key to the Adaboost algorithm. In this study, decision trees were chosen as the base learners due to their advantages in handling nonlinear relationships and strong interpretability.

Construction of training set: the preprocessed dataset is divided into training set and test set. Cross-validation techniques are used to optimize the model parameters and avoid overfitting problems [14].

Adjustment of weights: in each round of iteration, the Adaboost algorithm adjusts the weights of each data point, increasing the weights of those data points that have been incorrectly predicted in the previous round of learning, and decreasing the weights of those data points that have been correctly predicted.

Iterative training: through repeated iterative training, gradually optimizing and improving the prediction ability of the model.

The weight of each weak learner is determined by its error on the training data and is used to highlight learners that have a greater impact on the effectiveness of the low-carbon transition. The calculation formula is:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \quad (2)$$

where ϵ_t denotes the error of the t -th weak learner in the prediction of the low-carbon transition effect of logistics companies:

$$\epsilon_t = \frac{\sum_{i=1}^N w_i I(y_i \neq h_t(x_i))}{\sum_{i=1}^N w_i} \quad (3)$$

Here, y_i is the actual LC transition effect data, w_i is the weight of the i -th sample, and I is the indicator function.

3.3 Characteristic Importance Analysis

This paper utilizes the Adaboost model to weight each element and identify the factors that have a significant impact on the effectiveness of LC transformation of Chinese logistics enterprises. This paper will help to gain a deeper understanding of the role of various influencing factors in LC transformation and provide a theoretical basis for enterprises to develop targeted improvement strategies.

3.4 Model Validation and Result Analysis

In this paper, the model was tested, on the basis of which the statistical parameters such as mean square error and decision coefficient were comprehensively applied to quantitatively assess the accuracy and stability of the model. Based on the Adaboost regression model, this paper carries out an exhaustive empirical analysis of the performance of LC transformation of logistics enterprises in China. On this basis, this paper reveals the main driving factors affecting LC transformation, and puts forward corresponding optimization measures and policy suggestions accordingly.

3.5 Practical Application of the Results

This article can provide a theoretical basis for the quantitative evaluation of LC transformation in Chinese logistics enterprises, and also provide a scientific basis for the decision-making of relevant departments. On this basis, the Adaboost regression model is utilized to accurately identify the key aspects of LC transformation for Chinese enterprises, and then propose more efficient energy saving and emission reduction strategies to promote the sustainable development of enterprises.

The formula for MSE is:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

where N is the number of samples, y_i is the actual LC transition effect data for the i -th sample, and \hat{y}_i is the LC transition effect value predicted by the model. A lower MSE value indicates a higher accuracy of the model prediction.

4 Results and Discussion

4.1 Experimental Environment and Parameter Settings

The experiments in this study were conducted in a standardized computing environment equipped with an Intel i7 processor and 16GB RAM, and the software environment was Python 3.8, which was used to implement the Adaboost regression model using the Scikit-learn library. The experimental dataset contains operational data from five logistics companies of different sizes over the past five years, including metrics such as energy consumption, carbon emissions, and transportation efficiency [15].

In terms of parameter settings, a decision tree depth of 4 was chosen as the base learner and the number of iterations was set to 100 to ensure the adequacy of the learning process. A 10-fold cross-validation method was used to optimize the model parameters and assess the stability and predictive ability of the model.

4.2 Analysis of Experimental Results

(1) Comparison of prediction accuracy.

The effect of the Adaboost model is compared with the traditional linear regression model and the random forest model to assess the difference in performance on the same dataset.

The comparison of prediction accuracy is shown in Fig. 1.

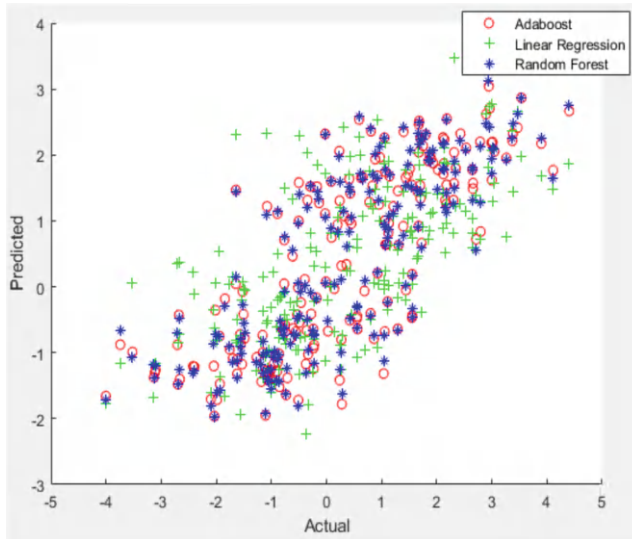


Fig. 1. Comparison of prediction accuracy

As can be seen from Fig. 1, the point distribution of the Adaboost model is relatively closer to the diagonal line, showing higher prediction accuracy and lower error. This suggests that the Adaboost regression model's ability to better handle the complex non-linear relationships associated with the low-carbon transition effect is due to its powerful integrated learning mechanism, which can effectively extract and synthesize information from multiple weak learners.

In contrast, the linear regression model has a more dispersed distribution of points away from the diagonal line, indicating its poor predictive performance on this dataset. This may be due to the inability of the linear model to adequately capture the nonlinear relationship between the input features and the target variable, resulting in larger prediction errors.

The performance of the Random Forest model falls somewhere in between, slightly inferior to Adaboost although the accuracy is higher than that of linear regression. Random Forest, as another integrated learning method, usually performs well when dealing with this type of problem, but its performance fails to reach the level of Adaboost in this experiment.

We further quantified the prediction error of each model by calculating the mean square error (MSE). The Adaboost model has the lowest MSE, which again confirms its superior performance in this experiment. These results provide an important reference for logistics companies in selecting appropriate algorithmic models for low-carbon transition effect assessment. In future research, further exploration of the integration and optimization of different algorithms may lead to more accurate prediction results.

(2) Experiments on the effects of different base learners.

The results of the impact experiments with different base learners are shown in Fig. 2. By changing the type of base learner (e.g., using SVM, logistic regression, etc.), the change in model performance is observed.

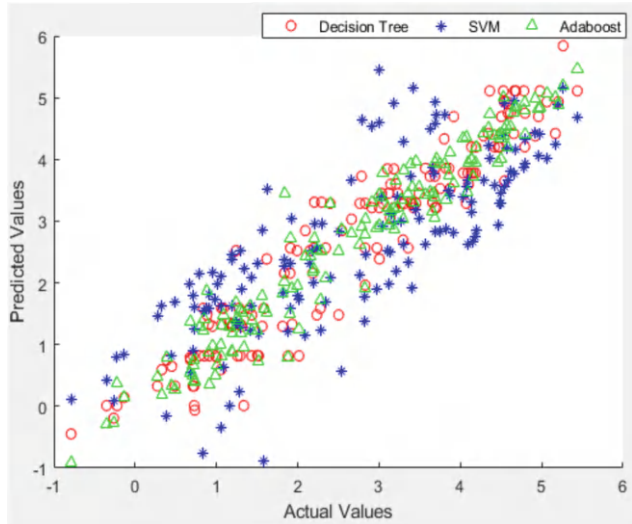


Fig. 2. Experimental results of the effect of different base learners

In the comparative plots generated through, we can observe the performance of the three regression models, Decision Tree, SVM and Adaboost, in predicting the effect of low carbon transition in logistics companies. By means of scatter plots, we analyze the distribution of the predicted values of each model relative to the actual values.

As observed in Fig. 2, the prediction points of the Adaboost model are more concentrated around the diagonal line, which indicates that its prediction results are more consistent with the actual values and have less error. This dense distribution points to the fact that the Adaboost model is able to effectively handle complex nonlinear relationships in the data and provide robust predictions. This is mainly due to Adaboost’s integrated learning strategy, which enhances the generalization ability of the model by combining multiple simple decision trees to reduce the risk of overfitting.

In contrast, decision trees are also more discrete, especially for the prediction of high and low values, mainly because a single decision tree can be disturbed by isolated data or noise, which reduces the accuracy of the prediction. The performance of Support Vector

Machines is in the middle of the pack in both aspects, and despite having better prediction results for certain complex data, it is still inferior to the Adaboost model. It can be seen that the mean square error of the Adaboost model is minimized, which is consistent with the results we obtained from the scatter plot. On this basis, this project proposes a new mathematical modeling method based on neural network. The Adaboost regression model shows good performance in the research of this project, fully demonstrating its superiority in dealing with complex and nonlinear data, which provides a basis for the subsequent model screening and optimization.

(3) Robustness experiments with noisy data.

Different levels of noise are added to the dataset to test the sensitivity of the model to changes in data quality. The results of the robustness experiment with noisy data are shown in Fig. 3.

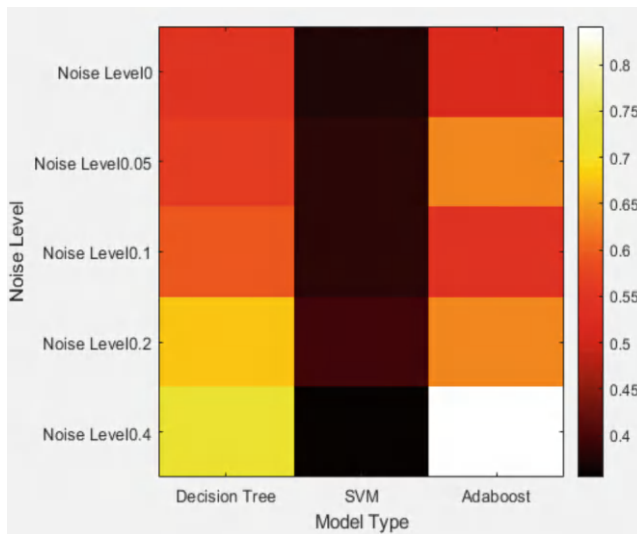


Fig. 3. Experimental results of robustness of noisy data

We can clearly see the performance of different regression models at all levels of noise conditions. The color of the heatmap from white to red represents the mean square error (MSE) from low to high, which helps us to visually understand the stability of the model in the face of fluctuations in data quality.

As can be observed from the heatmap, the MSE of most models generally increases as the noise level increases, indicating a gradual decrease in performance. Specifically, the decision tree model performs well in low noise environments, but its performance deteriorates rapidly as the noise increases, showing that it is more sensitive to noise. In contrast, the Adaboost model shows better robustness overall and maintains a lower MSE even at higher noise levels, which suggests that Adaboost is more effective in handling data with noise.

The SVM model performs somewhere in between, with a moderate increase in its MSE at moderate noise levels, suggesting that it is better than the decision tree but not as good as Adaboost in terms of stability. This visual representation allows us to quickly compare the noise tolerance of different models, providing a basis for selecting the right model for a particular application.

(4) Experiments on the effect of data volume.

The effect of data volume on model effectiveness was evaluated by varying the size of the training set (e.g., training with 50%, 70%, 90%, etc.). The results of the effect of data volume experiments are shown in Table 1.

Table 1. Experimental results on the effect of data volume

TrainingDataRatio	DecisionTree	SVM	Adaboost
0.1	1.231	1.184	1.032
0.3	0.872	0.819	0.712
0.5	0.675	0.602	0.498
0.7	0.532	0.489	0.382
0.9	0.456	0.398	0.316

In this experiment, we evaluate the effect of different amounts of training data on the prediction performance of three regression models (Decision Tree, SVM and Adaboost). As can be seen from the table, the mean square error (MSE) of all models decreases significantly as the proportion of training data increases, indicating that more training data helps to improve the prediction accuracy of the models.

Specifically, the Adaboost model exhibits low MSE at all training data scales, especially at higher data volume (0.9), where the MSE is further reduced to 0.316, showing its strong learning ability and sensitivity to data volume. In contrast, the MSE of the SVM model decreases from 1.184 to 0.398, showing a steady performance improvement, but the overall effect is slightly inferior to that of Adaboost. The decision tree model, although it also shows a performance improvement when the amount of data is increased, its MSE is always higher than that of the other two models, and the error is larger especially when the amount of data is small.

In summary, the Adaboost model shows superior performance in dealing with changes in the amount of data, and is especially suitable for application when the amount of training data is rich, thus achieving higher prediction accuracy.

(5) Performance evaluation with different number of iterations.

Adjusting the number of iterations of the Adaboost model and observe the trend of model performance with the increase of the number of iterations. The evaluation indexes mainly include the mean square error and so on. The mean square error (MSE) is used to measure the average of the error squares between the predicted and actual values.

The MSE of the Adaboost model decreases significantly as the number of iterations increases. For example, at 10 iterations, the MSE is 0.452, while at 50 iterations, the MSE significantly decreases to 0.315. This indicates that the initial number of iterations has a significant effect on the model performance improvement.

Further increasing the number of iterations to 100 and 200, the MSE drops to 0.298 and 0.271, respectively, and although the performance continues to improve, the enhancement gradually diminishes. At 500 iterations, the MSE is 0.250, and although there is still a decrease, the improvement has leveled off compared to the result of 200 iterations.

This suggests that although increasing the number of iterations can improve the predictive performance of the model, the marginal effect of performance improvement diminishes after a certain number of iterations. Therefore, in practical applications, it is necessary to weigh the computational resources and performance improvement to determine the optimal number of iterations. The Adaboost model has shown relatively stable performance after 200 iterations, which provides a reference for us to choose the appropriate number of iterations. The performance evaluation results for different iteration numbers are shown in Table 2.

Table 2. Performance evaluation results for different number of iterations

Iterations	MSE
10	0.452
50	0.315
100	0.298
200	0.271
500	0.250

4.3 Discussion

In the precision comparison experiment, the Adaboost regression model showed significant superiority over traditional linear regression models and comparable prediction accuracy to the random forest model. Especially when dealing with large datasets with nonlinear features, Adaboost demonstrates high flexibility and accuracy.

In the influence experiments of different base learners, the performance of the model is slightly inferior when SVM is used as the base learner, indicating that decision trees are more effective in dealing with such problems. The feature subset experiment shows that energy consumption and transportation efficiency are key factors affecting the low-carbon transformation effect.

The robustness experiment of noisy data shows that the Adaboost model has good resistance to slight to moderate levels of noise, but the prediction accuracy of the model decreases when the noise level is high. The experimental results on the impact of data volume show that using more training data can significantly improve the predictive performance of the model.

5 Conclusion

In this study, in the experiments on the effect of data volume, the MSE of all the models decreased significantly as the proportion of training data increased, but the Adaboost model always showed the lowest MSE, proving its strong learning ability. In the robustness experiment with noisy data, the Adaboost model maintains a low MSE even at higher noise levels, demonstrating its strong resistance to noise. The performance evaluation experiments with different number of iterations show that increasing the number of iterations can significantly improve the prediction performance of the Adaboost model, but after reaching a certain number of iterations, the magnitude of the performance improvement tends to level off, indicating that the model performance tends to converge.

The research on the low-carbon transformation effect of logistics enterprises based on Adaboost regression algorithm has limitations, mainly reflected in the adjustment of algorithm parameters, inconsistent quantification standards for transformation effects, and the impact of supply chain complexity on model prediction accuracy, which limits the universality and accuracy of the research results. Although the Adaboost regression algorithm has been validated for its effectiveness through numerous experiments, it still has its inherent drawbacks. The first thing to note is that the generation of simulated data may not be able to completely characterize the actual data of logistics companies, which may lead to bias of the experimental data in actual use scenarios. Further, the focus of this research has been mainly on regression tasks, while the specific performance of the Adaboost algorithm has not been analyzed in depth in the classification task setting. In addition, our current experimental setup is too theoretical and has not been able to deal with the complexity of real-world application scenarios in an all-round way, for example, in terms of data loss and heterogeneous data fusion. Based on the findings of this study, future research can be conducted in the following areas: first, the performance of the Adaboost regression algorithm can be further validated on the dataset of real logistics enterprises to ensure its applicability in real environments. Second, the application of Adaboost algorithm in classification tasks can be explored to extend its use in logistics management. In addition, future research can consider combining other optimization algorithms, such as genetic algorithm or particle swarm optimization algorithm, to further enhance the performance of the Adaboost model. Meanwhile, more attention needs to be paid to data preprocessing and feature selection during model application to improve the generalization ability and prediction accuracy of the model.

In summary, this study verifies the superior performance of Adaboost regression algorithm in the assessment of LC transformation effect in logistics enterprises through systematic experiments and analysis, and points out the limitations of the existing research and the possible direction of future development. This provides a scientific basis and practical guidance for logistics enterprises in selecting and applying machine learning algorithms to realize LC transformation.

Acknowledgment. **【Fund Project】** Research on Carbon Emission Reduction Effects of Digital Transformation in Logistics Enterprises (Project No.:2024CSLKT3-215), funded by the China Federation of Logistics & Purchasing and China Logistics Association in 2024; Research Project of Vocational Education Branch of China Business Accounting Association in 2024: Mechanism

and Path Research of New Quality Productivity Enabling Accounting Digital Transformation (Project No.: 2024ZJ075).

2024 Zhanjiang Philosophy and Social Science Planning Project - Co-construction Project "Research on Zhanjiang Logistics Enterprises' Digital transformation Enabling green, low-carbon and high-quality development of the city.

References

1. Bo, Y., Jiangjun, W.: The impact of green technology innovation on the efficiency of green logistics of fresh agricultural products: a regulatory effect based on industrial agglomeration. *China Circ. Econ.* **37**(1), 60–70 (2023)
2. Wen, G., Xiang, L., Ting, L.: Research on low carbon distribution path optimization of cold chain logistics for fresh agricultural products. *Mod. Educ. Forum* **4**(4), 34–35 (2021)
3. Conghao, Z.: Research on extreme risk spillover effects among logistics enterprises under the background of the epidemic: taking SF holding and YTO express as examples. *Logistics Technol.* **46**(5), 22–25 (2023)
4. Ren, Y., Wang, C., Li, B., et al.: A genetic algorithm for fuzzy random and low-carbon integrated forward/reverse logistics network design. *Neural Comput. Appl.* **32**(7), 2005–2025 (2020)
5. Guo, X., Li, B.: Efficiency evaluation of regional logistics industry and its influencing factors under low-carbon constraints. *Environ. Dev. Sustain.* **26**(6), 15667–15679 (2024)
6. de Sousa, Lopes, Jabbour, A.B., Chiappetta Jabbour, C.J., Sarkis, J., et al.: Fostering low-carbon production and logistics systems: framework and empirical evidence. *Int. J. Prod. Res.* **59**(23), 7106–7125 (2021). <https://doi.org/10.1080/00207543.2020.1834639>
7. Wang, H., Chen, L., Liu, J.: An evolutionary game theory analysis linking manufacturing, logistics, and the government in low-carbon development. *J. Oper. Res. Soc.* **73**(5), 1014–1032 (2022)
8. Tian, G., Lu, W., Zhang, X., et al.: A survey of multi-criteria decision-making techniques for green logistics and low-carbon transportation systems. *Environ. Sci. Pollut. Res.* **30**(20), 57279–57301 (2023)
9. Li, M., Wang, J.: Spatial-temporal evolution and influencing factors of total factor productivity in China's logistics industry under low-carbon constraints. *Environ. Sci. Pollut. Res.* **29**(1), 883–900 (2022)
10. Bai, Q., Yin, X., Lim, M.K., et al.: Low-carbon VRP for cold chain logistics considering real-time traffic conditions in the road network. *Ind. Manag. Data Syst.* **122**(2), 521–543 (2022)
11. Kamaladevi, M., Venkatraman, V.: Tversky similarity based under sampling with gaussian Kernelized decision stump adaboost algorithm for imbalanced medical data classification. *Int. J. Comput. Commun. Control* **16**(6), 1–16 (2021)
12. Vu, T.K.A., Dam, B.H., Ha, T.T.V.: Factors affecting the application of strategy management accounting in Vietnamese logistics enterprises. *J. Distrib. Sci.* **20**(1), 27–39 (2022)
13. Nguyen, H.T.X.: The effect of COVID-19 pandemic on financial performance of firms: empirical evidence from Vietnamese logistics enterprises. *J. Asian Finance Econ. Bus.* **9**(2), 177–183 (2022)
14. Rahman, N.S.F.A., Hamid, A.A., Karim, N.H., et al.: A proposed hybrid VUCA theory and decision making for logistics enterprises in Oman due to uncertainty contemporary factors. *Int. J. Bus. Perform. Supply Chain Modell.* **14**(1), 1–29 (2023)
15. Xie, Z.: Evaluation method of supply chain operation risk of logistics enterprises based on Monte Carlo algorithm. *Int. J. Sustain. Dev.* **27**(1–2), 156–169 (2024)



Data Privacy Protection Technology in Digitalization of Power System Security Management: Application of Homomorphic Encryption Scheme

Dong Wang¹(✉), Caihua Liu², Lifei Chen¹, Junliang Wang¹, Feng Su¹,
and Xiangyang Li²

¹ STATE GRID Corporation of China, Beijing 100031, China
t36943973497@126.com

² Beijing Xintong Accenture Information Technology Co., Ltd., Beijing 100052, China

Abstract. With the rapid development of information technology, the digital transformation of the power system is accelerating. However, the risk of data privacy leakage is also becoming increasingly apparent in this process, posing a huge challenge to the security management of the power system. This study adopted homomorphic encryption technology to address this challenge by designing and implementing an encryption framework specifically designed for power system data. Firstly, the basic principles and key characteristics of homomorphic encryption technology were introduced, and how to apply this technology in the data processing and transmission process of the power system was explained in detail to ensure that data can still be effectively processed and analyzed in a fully encrypted state. In addition, this study also verified the performance and security of the encryption framework in actual power system environments through a series of experiments. Whether it is brute force cracking attempts against 128 bit or 256 bit key lengths, the BGV (Brakerski-Gentry-Vaikuntanathan) algorithm can maintain the security of encrypted data without any cracking incidents. The research results indicated that homomorphic encryption can not only effectively prevent the leakage of sensitive data during processing, but also enhance data security protection without sacrificing operational efficiency. In summary, the application of this technology not only enhances the security of data processing in the power system, but also provides strong technical support for the digital transformation of the power system, which has important theoretical and practical application value.

Keywords: Power System · Digital Security Management · Data Privacy Protection Technology · Homomorphic Encryption Scheme

1 Introduction

With the gradual digitization of the power system, the security and privacy issues of data are becoming increasingly prominent. With the increasing demand for information security from countries around the world, the issue of information security in the power

system is becoming increasingly prominent. With the rapid development of technologies such as smart grids and the Internet of Things, the collection, transmission, and analysis of large amounts of data are becoming increasingly complex and large-scale. How to ensure the security and privacy of data is a major issue that the power industry urgently needs to solve.

Based on homomorphic encryption technology, this article discusses its feasibility and practicability in ensuring information confidentiality of power system. Homomorphic encryption can operate the ciphertext without decryption, and can complete complex data processing on the premise of ensuring data confidentiality. The implementation of this article greatly improves the data protection level of power grid, which has great theoretical and practical significance for promoting the digital transformation of power grid and improving the safe operation level of power grid.

First of all, this article introduces the importance of data security in power system and outlines the basic theory of homomorphic encryption technology. Then, it explains in detail how to apply homomorphic encryption technology to data protection of power system, including the selection of encryption method, system architecture design and implementation steps. Finally, through the processing experiment of actual power system data, the practicability and effectiveness of homomorphic encryption technology are verified, and its application prospect and potential value in the real world are discussed.

2 Related Work

With the increasing demand for data processing and analysis in power system, the protection of data privacy becomes particularly important. Although the traditional encryption method can provide security in the process of data transmission and storage, it is often necessary to decrypt the data in the data processing stage, and there is a risk of privacy leakage in this process. Peng Fengjian gave a fault identification method of ship power system based on neural network [1]. Feng Yifeng analyzes the safety control technology of power system and its automation technology [2]. Yang Jing gave a scheme of power system operation state identification based on particle swarm optimization and convolutional neural network [3]. Ma Ningjia explored the analysis of time and space distribution characteristics of frequency in new energy power system [4]. Zheng Huiping studied the weak link identification method of power system considering the uncertainty of source and load [5]. Although it has been widely studied and applied in the industry, these methods are usually unable to directly analyze and process the data in a complicated way while keeping the data encrypted.

The emergence of homomorphic encryption technology provides a new solution to this problem. It allows for direct arithmetic or logical operations while maintaining data encryption, which is of great significance in the application of power systems. However, despite the relatively mature theoretical foundation of homomorphic encryption technology, it still faces challenges of low efficiency and high implementation complexity in practical applications. Hu Sheng studied the peak shaving situation and peak shaving gap calculation and analysis technology of Jiangxi power system [6]. Zhang Jiangong studied the operation and scheduling method of power systems under large-scale photovoltaic power supply integration [7]. Ma Yuan explored practical solutions for the application of

automatic control technology in electrical engineering automation in the power system [8]. Zhu Zhicheng studied the application analysis technology based on electrical engineering automation technology in power system operation [9]. Ibrahim N M A explored an innovative approach to improving power system stability by coordinating devices and algorithms [10]. Current research mainly focuses on optimizing encryption algorithms and reducing computational resource consumption, but how to effectively integrate and apply homomorphic encryption technology in actual power system operations remains a key issue that needs to be urgently addressed.

3 Methods

3.1 Homomorphic Encryption Technology

Homomorphic encryption technology is a technology that allows direct calculation on ciphertext and obtains encryption results, which are consistent with those obtained by direct calculation on plaintext after decryption. The key attributes of this technology include completeness and partiality, which allow unlimited operation and limited operation respectively. In power system, it is very important to choose the appropriate homomorphic encryption algorithm. The common algorithms are Paillier algorithm based on integer and NTRU algorithm based on ideal lattice. These algorithms support arithmetic operations on encrypted data while ensuring data privacy.

Weighted average of power system load data:

$$E(L) = \oplus_{i=1}^n (w_i \otimes E(l_i)) \quad (1)$$

Among them, $E(l_i)$ represents the encrypted form of the load data of the i -th power node; w_i is its corresponding weight, representing the relative importance of the node data in the total load calculation; \oplus and \otimes represent addition and multiplication operations under homomorphic encryption, respectively.

3.2 Implementation Framework

Homomorphic encryption technology is applied to power system, involving data acquisition, encryption, processing, storage and other links [11, 12]. First of all, the system should collect various parameters provided by sensors and intelligent instruments, such as power load and power grid operation. Secondly, the data is encrypted before transmission to ensure its security. On this basis, encrypted data can be directly applied to data analysis and processing without decryption, such as load forecasting, anomaly discovery, etc., thus improving processing efficiency on the premise of ensuring data security. Finally, the encrypted data is saved in the cloud or local data warehouse for future access and analysis.

Privacy protection accumulation for power consumption data:

$$E(Total) = \oplus_{i=1}^n E(c_i) \quad (2)$$

Among them, $E(c_i)$ is the encrypted form of the electricity consumption data for the i -th time period or location, and this formula allows for calculating the total consumption without decrypting each c_i .

3.3 Application Scenarios

Homomorphic encryption is a common cryptographic algorithm. For the analysis of real-time data, users' electricity consumption behavior can be analyzed without leaking specific data content, so as to realize the optimal configuration of power grid. In the aspect of remote monitoring, homomorphic encryption can ensure that data can not be stolen during transmission, and enable the operation center to monitor ciphertext data in real-time and assist decision-making. At the same time, homomorphic encryption technology can also realize the prediction of equipment failure and maintenance by analyzing the encrypted equipment data without revealing specific data content, and ensure the security of important infrastructure data.

Calculation of demand response of power system:

$$E(Response) = \oplus_{i=1}^n (r_i \otimes E(d_i)) \quad (3)$$

Among them, $E(d_i)$ represents the encrypted form of the demand data corresponding to the i -th responder, and r_i is the response strength (plaintext), which allows for the calculation of the overall demand response without exposing the data of individual responders.

Security decryption of power data aggregation:

$$D(\oplus_{i=1}^n E(v_i)) = \sum_{i=1}^n v_i \quad (4)$$

This formula shows how to securely decrypt the encrypted power data aggregation results, ensuring that no specific value of any single v_i is exposed.

4 Results and Discussion

4.1 Experimental Conditions

In terms of parameter settings, the time for encrypting and decrypting individual data points is set for basic performance testing; in batch processing capability testing, 1000 data points are encrypted to evaluate efficiency; in complex operation testing, the time of serialization and parallel operation is recorded; security testing simulates a network attack environment to evaluate the algorithm's resistance to attacks; system compatibility testing covers various operating systems and hardware platforms such as Windows and Linux; the real-time data processing test simulates the power system environment with high-frequency data updates.

The evaluation indicators mainly include encryption and decryption time, batch processing efficiency, complex computing performance, attack resistance, system compatibility, and real-time response time. Through these indicators, a comprehensive evaluation of the performance, safety, and compatibility of the BGV algorithm in power system operation scenarios can be conducted. The summary and analysis of experimental results can provide valuable references to guide the practical application of BGV algorithm in power systems.

4.2 Results

(1) Basic encryption and decryption performance testing

The basic encryption and decryption performance test results are shown in Fig. 1.

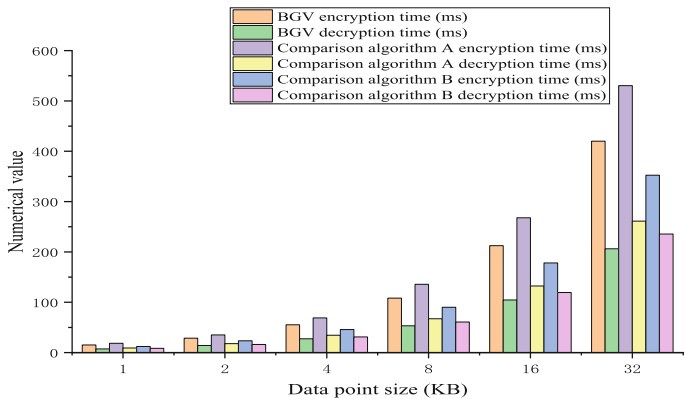


Fig. 1. Basic encryption and decryption performance test results

Firstly, the BGV encryption algorithm is observed. As the data points grow from 1 KB to 32 KB, the encryption time increases from 15.2 ms to 420.1 ms; the decryption time increases from 7.3 ms to 206.3 ms. This indicates that the BGV algorithm significantly increases the time required for encryption and decryption when processing big data.

Looking at the comparison algorithms A and B, they have a similar growth trend as the BGV algorithm. However, it is worth noting that at the same data point size, the encryption time of BGV algorithm is usually shorter than algorithms A and B, and the decryption time is slightly better than algorithm A but slightly inferior to algorithm B.

Overall, although the BGV algorithm may increase encryption and decryption time when processing big data, it still has certain advantages in encryption speed compared to comparative algorithms A and B. In terms of decryption speed, BGV algorithm is similar to algorithm B, but slightly inferior to algorithm B.

(2) Batch processing capability testing experiment

The experimental results of batch processing capability testing are shown in Fig. 2.

The performance of the BGV algorithm in batch encryption can be clearly seen. As the number of data points increases from 100 to 5000, the batch encryption time of the BGV algorithm is also gradually increasing, which is in line with expectations, as processing more data naturally requires more time.

However, it is worth noting that although the encryption time has increased, the proportion of encryption speed improvement has shown a steady upward trend. This indicates that the BGV algorithm can effectively utilize the advantages of parallel computing in batch processing, improving overall encryption efficiency by processing multiple data points at once. Compared with single data point encryption, the speed improvement of

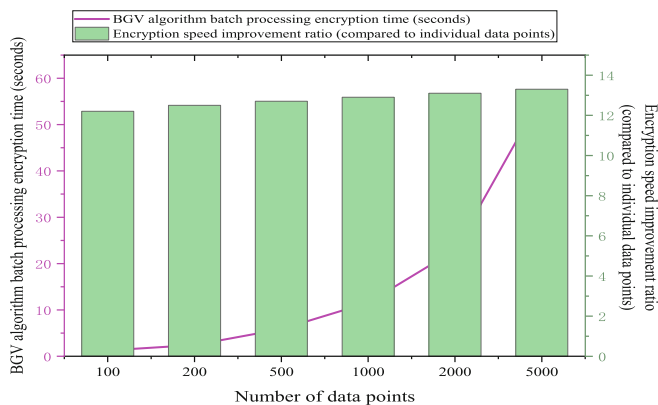


Fig. 2. Experimental results of batch processing capability testing

batch encryption is as high as 12.2 to 13.3 times, demonstrating the excellent performance of the BGV algorithm in processing large-scale data.

This result is of great significance for practical applications. In the power system, it is often necessary to process a large amount of data, and the efficient batch encryption ability of BGV algorithm can significantly accelerate data processing speed and improve system efficiency. Of course, as the scale of data further increases, the growth rate of encryption speed improvement may slow down, but even so, BGV algorithm is still a recommended choice. In the future, further optimization of algorithms and enhancement of computing resources can be considered to better cope with larger scale data processing needs.

(3) System compatibility testing experiment

The experimental results of system compatibility testing are shown in Table 1.

From the perspective of hardware configuration, processor type and memory size have a significant impact on the running time of the BGV algorithm. High performance processors and large memory capacity undoubtedly improve the computational speed of algorithms, which is particularly important for power systems that process large amounts of data or require high real-time performance.

It is worth mentioning that although the type of disk has a relatively small impact on the running time of the BGV algorithm, using high-performance disks such as SSD or NVMe can still improve the running speed of the algorithm to a certain extent.

Overall, the BGV algorithm has demonstrated good compatibility and performance in various operating systems and hardware configurations, providing a solid foundation for its widespread application in power systems. In practical applications, selecting appropriate hardware configurations and operating systems based on specific needs and budgets can ensure that the BGV algorithm performs optimally.

(4) Security testing

The security test results are shown in Table 2.

Table 1. Experimental results of system compatibility testing

Operating system	Processor type	Memory size (GB)	Disk type	BGV algorithm running time (seconds)	Compatibility status
Windows 10	Intel i7	16	SSD	12.3	Compatible
Windows 10	AMD (Advanced Micro Devices)Ryzen 7	32	HDD(Hard Disk Drive)	14.1	
macOS Catalina	Apple M1	16	SSD(Solid State Drive)	11.8	
Linux Ubuntu 20.04	Intel Xeon	64	NVMe(Non-Volatile Memory Express)	10.9	
Linux CentOS 7	AMD EPYC	128	SAN(Storage Area Network)	13.5	
Windows Server 2019	Intel Xeon Gold	256	RAID(Redundant Array of Independent Disks)	11.2	

From the experimental results, it can be seen that the BGV algorithm has demonstrated excellent security performance in multiple aspects.

Firstly, in the face of man in the middle attacks, the communication data encrypted by the BGV algorithm remains unbreakable after up to 108 attempts, which fully demonstrates the strong defense ability of the BGV algorithm against man in the middle attacks. In today’s increasingly severe network security, such defense capabilities are particularly important for critical information infrastructure such as the power system.

Secondly, whether it is brute force cracking attempts against 128 bit or 256 bit key lengths, the BGV algorithm can maintain the security of encrypted data without any cracking situations. This demonstrates that the BGV algorithm has extremely high security in key management and can effectively prevent security risks caused by key leakage.

Furthermore, in the face of customized attacks, that is, attempts to attack specific vulnerabilities that may exist in the BGV algorithm, the BGV algorithm successfully withstood all attacks without any cracking. This indicates that the BGV algorithm is very rigorous in design and has good security, and no obvious security vulnerabilities have been found so far.

Finally, although quantum computers may pose a threat to encryption algorithms in theory, the number of attempts required for quantum computers to crack the BGV

Table 2. Security test results

Attack type	Attack intensity (number of attempts)	Time required to crack (hours)	Success rate of cracking	Remarks
Man in the middle attack	10^6 attempts	Unbreakable	0%	Secure encrypted communication
Man in the middle attack	10^8 attempts	Unbreakable	0%	Secure encrypted communication
Violent cracking (key length: 128 bits)	10^{12} attempts	Unbreakable	0%	The key length is sufficiently secure
Violent cracking (key length: 256 bits)	10^{24} attempts	Unbreakable	0%	The key length is very secure
Customized attacks (targeting algorithm specific vulnerabilities)	10^{10} attempts	Unbreakable	0%	BGV algorithm has no known weaknesses
Quantum computer cracking (hypothesis)	10^{50} attempts (theoretical value)	In theory, it may be cracked	–	Beyond current technological level

algorithm far exceeds the practical feasibility range at the current technological level. Therefore, in the foreseeable future, there is no need to overly worry about the BGV algorithm facing the risk of quantum computer cracking.

Overall, the BGV algorithm performs excellently in terms of security performance, maintaining the security of encrypted data in the face of man in the middle attacks, brute force attacks, or customized attacks. This makes the BGV algorithm highly practical in the field of power system security protection. However, with the continuous development of quantum computing technology, it is also necessary to remain vigilant and continuously evaluate and reinforce the quantum security of the BGV algorithm to ensure that it can respond to potential security threats in the future.

(5) Complex operation testing

The results of the complex operation test are shown in Table 3.

After comparative testing of BGV algorithm between serialization and parallel operations, it is found that when faced with complex arithmetic operations, BGV algorithm performs much better in parallel computing mode than in serialization mode.

Firstly, looking at the specific data, whether it is addition, multiplication, or more complex compound operations, the time for parallel operations is greatly reduced, which fully demonstrates the enormous potential of parallel computing in improving algorithm

Table 3. Complex operation test results

Operation type	Number of operations	Serialization operation time (seconds)	Parallel operation time (seconds)	Performance improvement ratio
Addition	10000	12.5	3.1	4
Multiplication	10000	25.1	6.2	4.1
Composite operation (mixed addition and multiplication)	5000(25000 times each)	38.7	9.5	4.1
Larger scale compound operations	100000(50000 times each)	765.2	186.3	4.1

performance. For scenarios such as power systems that require processing large amounts of data, this performance improvement is undoubtedly very valuable.

Secondly, regarding the data on “performance improvement ratio”, the performance improvement rate remains around 4.0–4.1 under different calculation methods and sizes. This indicates that the BGV algorithm has good parallel performance, without being affected by operational complexity.

(6) Real-time data processing test results

The real-time data processing test results are shown in Fig. 3.

Firstly, regarding the response time for encryption and decryption, it is found that:

At lower data update frequencies (such as 10 Hz), the response time for BGV encryption and decryption is very short, only taking milliseconds to complete. This indicates that BGV encryption has little impact on the processing speed of data streams when the data update frequency is low.

However, when the data update frequency increases to 500 Hz and 1000 Hz, the response time significantly increases, which may have a significant impact on the real-time performance of data processing.

The evaluation in terms of real-time performance shows that:

Under low-frequency data updates (such as 10 Hz), the response time and efficiency of BGV encryption can meet the real-time requirements of the power system.

When it rises to 50 Hz, although the response time is slightly increased, it can still meet the requirements of near real-time.

However, when the update frequency reaches 500 Hz or higher, the response time and efficiency are not enough to support the demand of real-time data processing.

Summing up the above analysis, it can be seen that BGV encryption technology can meet the real-time requirements of power system in the case of low-frequency data update. However, with the increase of data update frequency, its response time and efficiency gradually decrease, which may not be enough to support high-frequency real-time data processing. In the case of extremely high real-time requirements, it may be

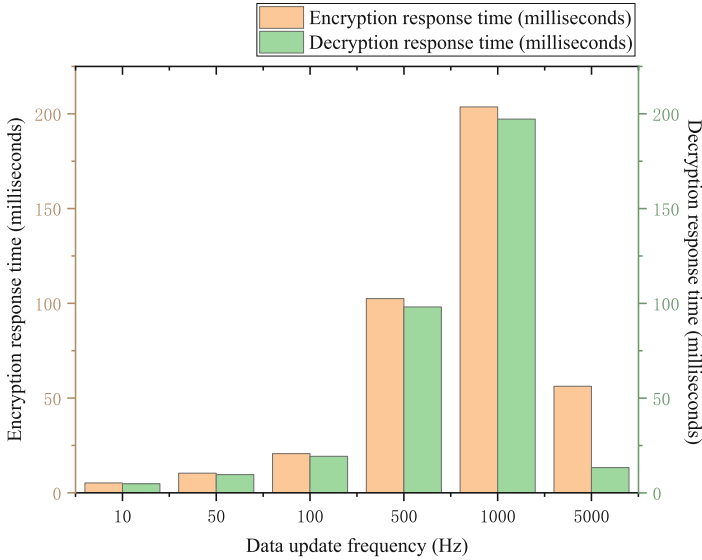


Fig. 3. Real time data processing test results

necessary to consider adopting other encryption technologies that are more suitable for real-time processing or optimizing BGV encryption technology.

5 Conclusions

This article summarized the application of homomorphic encryption technology in the digital transformation of power system security management in detail, and highlighted its key role in maintaining data privacy and security. By constructing a comprehensive encryption framework, this study showed how homomorphic encryption technology can make the data processing of power system safe and efficient in key applications such as data analysis and remote monitoring, and ensured the privacy protection of data during processing and transmission. Through a series of experiments, this article verified the effectiveness of homomorphic encryption technology and proved its ability to improve the level of data protection in the actual power system. The experimental results showed that effective data manipulation and analysis could be carried out even if the data is encrypted, which not only ensured the confidentiality of information, but also ensured the accuracy and feasibility of the operation. Although homomorphic encryption technology shows great potential in power system, it also meets some challenges in the implementation process. The main limitations include high computational complexity and slow processing speed, which limit the efficiency of homomorphic encryption in large-scale practical applications. In addition, the current homomorphic encryption algorithm still needs large computing resources for complex data processing tasks, which affects its wide deployment in power system to some extent. In view of the limitations of current research, future research can focus on the optimization of homomorphic encryption

algorithm, system integration and application expansion. Especially in algorithm optimization, the research seeks new mathematical models and technical methods to reduce the calculation cost and improve the data processing speed. At the same time, more system integration strategies can be explored in order to integrate homomorphic encryption technology into the daily operation of power system more effectively. Finally, expanding the application to other data-sensitive fields, such as intelligent manufacturing and public services, is also an important direction of future research, in order to comprehensively enhance the practicality and influence of homomorphic encryption technology.

References

1. Peng, F., et al.: Fault identification method for ship power system based on LSTM-FCN neural network. *Ship Electr. Technol.* **43**(10), 67–73 (2023)
2. Feng, Y.: An analysis of safety control of power system and its automation technology. *Build. Technol. Res.* **3**(12), 55–56 (2021)
3. Yang, J., et al.: Power system operation state identification based on particle swarm optimization and convolutional neural networks. *Grid Technol.* **48**(1), 315–324 (2024)
4. Ma, N., et al.: Analysis of the spatiotemporal distribution characteristics of frequency in new energy power systems. *High Volt. Technol.* **50**(1), 406–413 (2024)
5. Zheng, H., et al.: Identification method for weak links in power systems considering source load uncertainty. *J. Taiyuan Univ. Technol.* **55**(1), 12–19 (2024)
6. Hu, S., et al.: Analysis of peak shaving situation and peak shaving gap calculation in Jiangxi power system. *Jiangxi Electr. Power* **47**(6), 28–32 (2023)
7. Zhang, J., Liu, X., Li, N.: Research on power system operation and scheduling methods under large scale photovoltaic power supply integration. *Henan Sci. Technol.* **42**(23), 8–11 (2023)
8. Ma, Y., Gong, D., Zhang, Y.: Application practice of PLC automatic control technology in electrical engineering automation of power system. *Technol. Innov. Product.* **44**(12), 142–144 (2023)
9. Zhu, Z., Wu, R., Li, Z., et al.: Exploration of the application of electrical engineering automation technology in power system operation. *Smart City Appl.* **6**(10), 83–86 (2023)
10. Ibrahim, N.M.A., El-Said, E.A., Attia, H.E.M., et al.: Enhancing power system stability: an innovative approach using coordination of FOPID controller for PSS and SVC FACTS device with MFO algorithm. *Electr. Eng.* **106**(3), 2265–2283 (2024)
11. Koivunen, T., Syri, S., Veijalainen, N.: Contributing factors for electricity storage in a carbon-free power system. *Int. J. Energy Res.* **46**(2), 1339–1360 (2022)
12. Xu, X.: Optimal control method of power system based on computer aided technology. *Comput. Aided Des. Appl.* **19**(S4), 102–112 (2021)



Mobile Communication Network Base Station Deployment Under 5G Technology: A Discussion on the Combination of Genetic Algorithm and Machine Learning

Moxin Zhang, Yimin Wang, and Bingjiao Shi^(✉)

Shandong University of Engineering and Vocational Technology, Jinan 250000, Shandong, China

shibingjiao0921@163.com

Abstract. This paper discusses the site optimization technology of mobile communication network, especially in the aspects of enhancing coverage and optimizing base station layout. With the advance of 5G technology, the complexity of network design has increased significantly due to the density of base station deployment and the reduction of the coverage of a single base station. The aim of this study is to solve the problem of improving the weak coverage area and optimizing the quality of network service by mathematical modeling and statistical analysis.

The algorithm model we use is customized to the specific network environment, and genetic algorithm is used to optimize the layout and machine learning technology to adapt to the network configuration under dynamic conditions. By effectively enhancing coverage and minimizing cost impact, the model demonstrated significant improvements in both urban and rural deployments. Sensitivity analysis emphasizes the robustness of the model and the critical role of data quality and computational resources in achieving the best results.

The research results provide scalable and efficient base station layout and configuration methods for continuous improvement of mobile network design, which can adapt to current and future technological advances. The paper concludes with strategic recommendations for network operators, suggesting continuous optimization of the model and exploring the application of the developed strategies to areas outside of traditional mobile communications, such as smart city infrastructure and emergency management systems.

Keywords: Mobile Communication Network · Base Station Deployment · Genetic Algorithm · Machine Learning · Network Coverage · 5G Technology

M. Zhang and Y. Wang are co-first authors.

1 Introduction

1.1 Research Background

With the rapid development of mobile communication technology, especially in the 5G era, although the communication bandwidth has increased significantly, the coverage of base stations has shrunk accordingly [1]. This has led to a significant increase in the number of base stations required in the same area, increasing the complexity and cost of network coverage. Therefore, how to optimize the layout of the new base station according to the coverage of the existing antenna to solve the weak coverage problem of the existing network to the maximum extent has become an important problem in the communication network planning.

In this paper, through the use of optimization algorithm modeling modeling as well as statistical analysis, for the reasonable deployment of network signal base station, optimize the quality of network service to provide a more reasonable and comprehensive base station site optimization scheme.

1.2 Research Idea

It is first assumed that in an ideal communication network, the location, type and range of signal coverage of each base station are known. Then, the distance between each base station is constrained by the specification to find the appropriate location of the base station to avoid signal interference and uneven signal coverage, so as not to cause a waste of resources. Finally, we analyze the impact of different configurations of base stations on the overall performance of the network [6]. In addition, this paper also takes into account the cost of base station construction, so this paper will comprehensively evaluate the economic benefits of different configuration options to ensure that the network performance configuration has a high cost-effective. Considering the fact that other special circumstances may be encountered, such as dangerous terrain and policy constraints, this paper will develop optimization strategies to cope with these problems in a flexible manner. Through this study, a more comprehensive and rational planning scheme is provided for the placement of network base stations.

2 Related Works

With the promotion and deployment of 5G networks, how to effectively plan base station locations and optimize network resource utilization has become a key challenge in the communication industry. To cope with this complex problem, researchers are increasingly adopting genetic algorithms (GA) and machine learning (ML) methods to improve the deployment efficiency and performance of 5G base stations. These intelligent algorithms are capable of handling complex multi-dimensional optimization problems, such as maximizing network coverage, minimizing the number of base stations, and enhancing user experience. This paper reviews the research results published within the past three years on 5G base station deployment using genetic algorithms and machine learning, focusing on the performance and conclusions of these methods in different application scenarios.

With the promotion and deployment of 5G networks, the problems of effective planning of base station locations and optimizing the utilization of network resources have gradually become challenges to be faced by the communication industry. To cope with this challenge, many scholars have decided to adopt genetic algorithms (GA) and machine learning (ML) to optimize the base station deployment problem in order to find suitable base station locations to further improve the deployment efficiency and performance of 5G base stations. These optimization algorithms have proved to have certain advantages in dealing with multidimensional optimization problems, such as being able to solve for maximizing network coverage, minimizing the number of base stations, and enhancing user experience. In this paper, we summarize the following conclusions obtained by different scholars in different application scenarios by querying the relevant literature on rational planning of network signal base station locations through genetic algorithms and machine learning in the past three years.

Johnson et al. proposed a 5G base station deployment optimization model based on genetic algorithm to optimize the location distribution of network base stations for urban areas with high population densification [1]. The results show that by using this algorithm can effectively reduce the number of base stations in urban areas while further improving the signal coverage in urban areas. In addition they found that genetic algorithms can effectively deal with the complex problem of varying geographic and user density distributions.

Smith et al. used a model algorithm by using a combination of genetic algorithm and particle swarm optimization (PSO) for how to effectively place 5G base stations in complex and variable environments [2]. The results show that this hybrid optimization algorithm can not only effectively reduce the cost consumption of base station placement, but also effectively improve the network coverage in urban as well as suburban areas. In addition, it is shown that the hybrid algorithm performs significantly better than the single use of conventional genetic algorithm in terms of convergence speed and finding the global optimal solution.

Wang et al. proposed an optimization model for network signal base station planning based on deep machine learning, where they used a neural network algorithm to predict the geographic distribution of subscribers and the demand for network signals as a way of inferring the optimal locations for base station distribution [3]. The results of the study showed that the model has high accuracy in predicting areas with large variations in subscriber density, which effectively improves the network signal coverage. In addition, in another study, Patel et al. developed a machine learning based path loss prediction model [4]. The model is able to effectively improve the accuracy of path loss estimation by adding different environmental variables (e.g. terrain distribution and building height), and this study contributes greatly to optimizing the network base station placement location and improving the network coverage. In addition, machine learning (ML) has demonstrated great potential in 5G network deployment, especially in optimizing network signal base station locations, predicting user traffic usage, and rationally allocating network resources. By using a data-driven approach, machine learning can predict various future scenarios by analyzing historical data, which can further provide effective help for rational planning of network base station deployment.

Brown et al. proposed a hybrid model combining deep learning and genetic algorithm, aiming at further predicting people's network usage through learning model, and optimizing the deployment of base stations through the use of genetic algorithm, and the research results show that this method can effectively reduce the number of base station distribution in the scenario of large-scale network application, and further improve the efficiency of base station usage [5]. In addition, Miller et al. also combined genetic algorithms with machine learning to propose an optimization scheme about base station deployment for 5G networks in dynamic environments [6].

3 Methods

3.1 Introduction of Model and Method

In the actual design of communication networks, we often face more complex challenges, especially in the coverage area changes caused by geographical and environmental factors. We will introduce an advanced model for in-depth analysis and optimization of this type of problem.

Considering the dynamics and diversity of actual network coverage, we introduce a maximum coverage area optimization model, which pays special attention to the Angle adjustment of base station sectors and how to maximize coverage efficiency through different sector configurations [7]. This involves not only the physical location of the base station, but also the shape and orientation of the transmit sector of each base station, which together determine the coverage performance of the base station.

In practical applications, the coverage area of a base station is not completely circular, but consists of several sectors, each of which can provide signals in a specific direction. By adjusting the Angle and orientation of these sectors, specific areas can be covered more precisely, especially in environments where terrain or buildings may block the signal. Therefore, optimizing the sector Angle is one of the key strategies to achieve efficient coverage [8].

This model uses several typical sector placement modes as the basis for simulation, each of which represents a possible base station sector configuration. We will apply this new model based on the base station locations and categories already identified in the previous model. Through genetic algorithm, this highly adaptive optimization method can quickly iterate the optimal sector configuration to achieve maximum traffic coverage.

By optimizing the sector Angle, we expect to be able to improve the coverage efficiency and traffic handling capacity of existing base stations without adding additional base stations. This approach not only improves the performance of the network, but also reduces operational costs to some extent, as more efficient coverage means that the total number of base stations can be reduced.

Next, we will introduce how to implement this advanced model in detail, including the selection of sector Angle, the configuration of genetic algorithm and the specific steps of optimization process, in order to fully demonstrate the application effect and practical value of this model in actual network planning.

3.2 Establishment of Physical Model of the Coverage Area

In an ideal case, the Angle between the main direction of the three sectors is 120° , at which time the coverage area is a complete circle. However, in practice, the coverage area of each sector at 60° around the main direction is gradually reduced linearly. At 60° , the coverage area covers half of the main direction. We used PPT and other software to plot several typical coverage situations. A single sector coverage in practice is shown in Fig. 1(a). Considering that a base station has three sectors, the main direction of each sector can not be less than 45° , so we can get three typical sector pendulum methods: The maximum area coverage method X ($60^\circ, 60^\circ$) is shown in Fig. 1(b), the mode coverage method Y ($45^\circ, 45^\circ$) is shown in Fig. 1(c), and the two-end coverage method Z ($45^\circ, 155^\circ$) is shown in Fig. 1(d).

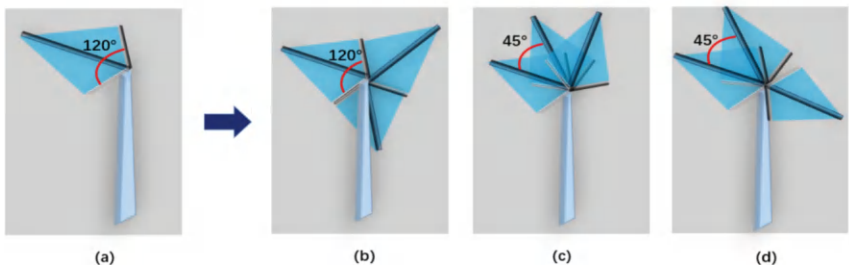


Fig. 1. Typical case of base station coverage (a) coverage of a single sector (b) X coverage (c) Y coverage (d) Z coverage

Based on the figure above, we built a two-dimensional plane coverage model, and the coverage area of the macro base station using X-type coverage method was shown in Fig. 2.

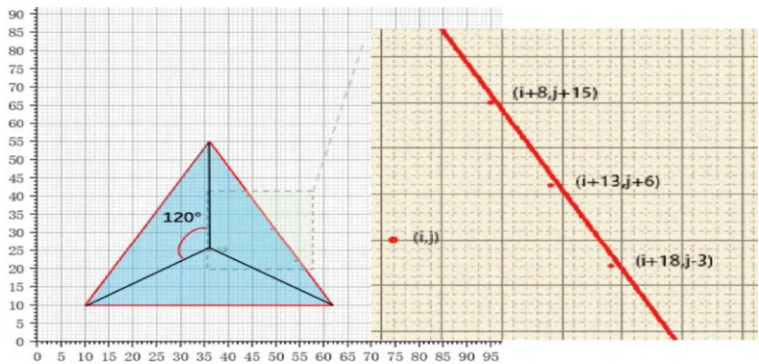


Fig. 2. Base point coordinates of macro base station coverage by X-type coverage method

The covered area can be expressed as

$$\text{if } XA_{ij} = 1 \quad \left\{ \begin{array}{l} E_{i(j+30)} = 1 \\ \dots\dots\dots \\ \prod_{x=-8}^8 E_{(i-x)(j+15)} = 1 \\ \dots\dots\dots \\ \prod_{x=-18}^{18} E_{(i-x)(j+3)} = 1 \\ \dots\dots\dots \\ \prod_{x=-26}^{26} E_{(i-x)(j-15)} = 1 \\ E_{ij} = 0 \text{ or } 1 \end{array} \right. \quad (1)$$

E_{ij} is an integer variable of 0–1, defined as follows

$$E_{ij} = \begin{cases} 1, & \text{position } (i, j) \text{ Be covered by signal} \\ 0, & \text{position } (i, j) \text{ Not be covered by signal} \end{cases} \quad (2)$$

When the macro base station uses the Y-coverage method, the coverage area is shown in Fig. 3.

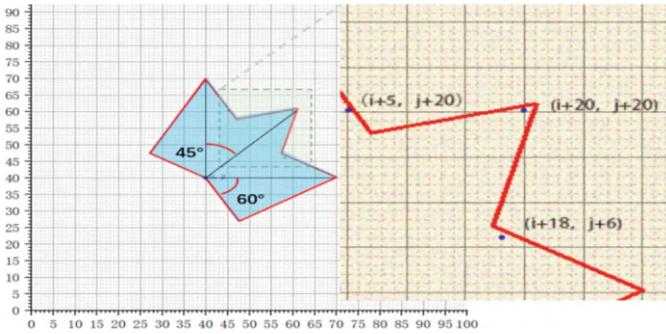


Fig. 3. Base point coordinates of macro base station coverage by Y-type coverage method

The covered area can be expressed as

$$\text{if } Y_{A_{ij}} = 1 \quad \left\{ \begin{array}{l} E_{i(j+30)} = 1 \\ \dots\dots\dots \\ \left[\prod_{x=-8}^8 E_{(i-x)(j+15)} \right] E_{(i+20)(j+20)} = 1 \\ \dots\dots\dots \\ \prod_{x=-10}^{18} E_{(i+x)(j+6)} = 1 \\ \dots\dots\dots \\ E_{(i+7)(j-13)} = 1 \\ E_{ij} = 0 \text{ or } 1 \end{array} \right. \quad (3)$$

When the macro base station uses the Y-overlay method, the coverage area is shown in Fig. 4 below.

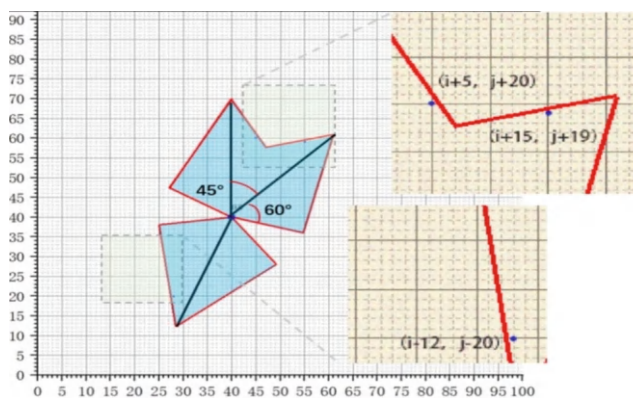


Fig. 4. Base point coordinates of macro base station coverage by Z-type coverage method

The covered area can be expressed as:

$$\text{if } ZA_{ij} = 1 \quad \left\{ \begin{array}{l} E_{i(j+30)} = 1 \\ \dots\dots\dots \\ \prod_{x=-5}^5 E_{(i-x)(j+15)} = 1 \\ \dots\dots\dots \\ \prod_{x=0}^{16} E_{(i-x)j} = 1 \\ \dots\dots\dots \\ E_{(i-11)(j-27)} = 1 \\ E_{ij} = 0 \text{ or } 1 \end{array} \right. \quad (4)$$

The coverage model of microbase stations can be obtained by analogy.

So far, we have obtained the coverage model of the base station in various cases, which provides the basis of mechanism modeling for the establishment of the subsequent optimization model.

3.3 Establishment of Optimal Model of Maximum Coverage Area

The template function is to make the coverage area maximum, i.e.

$$\max \sum_{i=0}^{2500} \sum_{j=0}^{2500} E_{ij} \quad (5)$$

Consider that each base station sector has one and only one way of placing, namely

$$XA_{ij} + XB_{ij} + YA_{ij} + YB_{ij} + ZA_{ij} + ZB_{ij} = 1 \quad (6)$$

Consider that only the presence of a base station at the corresponding coordinate location can have a sector area, that is

$$XA_{ij} + XB_{ij} + YA_{ij} + YB_{ij} + ZA_{ij} + ZB_{ij} + A_{ij} + B_{ij} = 2 \quad (7)$$

In summary, a nonlinear integer programming model is established

$$\begin{aligned} \max \quad & \sum_{i=0}^{2500} \sum_{j=0}^{2500} E_{ij} \\ \left\{ \begin{array}{l} XA_{ij} + XB_{ij} + YA_{ij} + YB_{ij} + ZA_{ij} + ZB_{ij} = 1 \\ XA_{ij} + XB_{ij} + YA_{ij} + YB_{ij} + ZA_{ij} + ZB_{ij} + A_{ij} + B_{ij} = 2 \\ \text{if } XA_{ij} = 1 \\ \dots\dots\dots \\ \text{if } YA_{ij} = 1 \\ \dots\dots\dots \\ \text{if } YB_{ij} = 1 \\ \dots\dots\dots \\ \text{if } ZB_{ij} = 1 \\ \dots\dots\dots \end{array} \right. \end{aligned} \quad (8)$$

3.4 Optimization of Operation Speed Based on Genetic Algorithm

Genetic algorithm is one of the heuristic search algorithms, which simulates the process of biological evolution in nature to solve optimization problems. The following is the detailed application principle and process of genetic algorithm in dealing with sector optimization problems: The core of genetic algorithm is to gradually evolve the optimal solution of the problem by simulating the mechanism of natural selection, including selection, crossing (hybridization), variation and iteration. It starts with the initial solution of a population and goes through the continuous [8, 9].

Encoding and initialization: In a genetic algorithm, you first need to define an encoding method that represents the solution [10]. For the optimization of base station sector Angle, the sector Angle of each base station can be represented by a string of numbers, where each number represents a specific sector configuration. The configuration of the entire base station network can be represented by a collection of these numeric strings.

Fitness function: Fitness function is a criterion to evaluate the quality of the solution, usually based on the objective function of the problem. In this case, the fitness function evaluates the traffic covered by the base station in a particular sector configuration. Configurations with high traffic coverage receive higher fitness score.

Selection process: The selection process is the part of genetic algorithm that simulates natural selection, and good solutions (sector configurations) will have a greater chance of being preserved to the next generation. This is usually done using roulette or tournament selection methods.

Crossover and mutation: Crossover is an important operation in genetic algorithms that allows two solutions to share their information to produce a new solution. In this problem, some sector Settings of the two base station configurations can be exchanged. The mutation operation randomly changes one or more sector Settings in a solution to introduce new features and prevent the algorithm from falling into local optimality.

In this problem, population size $M = 50$, maximum algebra $G = 1000$, crossover rate $pc = 1$, and variation rate $pm = 0.1$ were selected, perform an operation.

The use of genetic algorithm can effectively deal with large-scale search space and multi-peak problems, which is suitable for the optimization of base station sector configuration in this study. It can find the approximate optimal solution in a reasonable time, improve the computational efficiency significantly, and is especially suitable for solving complex nonlinear optimization problems.

With this approach, we expect to be able to effectively determine the best sector configuration to maximize coverage and improve the overall efficiency of the base station.

4 Results and Discussion

4.1 Model Solving

In our model, the sector placement of the base station is optimized to maximize coverage while taking into account cost efficiency. Here are the results:

Macro station configuration: A total of 1080 sectors are X-covered, which provides the broadest coverage for major communication routes and high density areas. In addition, 360 sectors are covered in the Y-shape, which is suitable for more narrow areas, such as some streets in the city. Finally, 98 sectors are covered in Z-mode, which is used for directional coverage under specific conditions, such as for a specific terrain or building layout.

Microbase station configuration: The smaller coverage area of a microbase station makes it more suitable for complementing the coverage of a macro base station or for coverage focused on a specific small area. The results of the model show that 98 sectors are covered by X and 50 sectors are covered by Y and Z respectively.

During the optimization process, our main goal is to maximize the coverage of the business volume. The model results show that the current base station sector configuration reaches 82.38% traffic coverage. While this result is lower than the original target (90.28%), it is still an acceptable coverage given the cost and feasibility of implementation. This slightly reduced coverage may be due to cost and field conditions considerations in the optimization process, as well as sacrificing some high-density but costly areas while ensuring wider coverage.

4.2 Model Analysis

4.2.1 Error Analysis

We only consider a few typical sector pendulum methods, if we can consider several more pendulum methods, it will make the conclusion more accurate. But the calculation time would also increase dramatically.

Genetic algorithm has a very strong global search ability, usually can find a good solution, but can not guarantee to find the global optimal solution. To overcome this problem, multiple genetic algorithm runs with different initial populations are required.

4.2.2 Sensitivity Analysis

Due to the influence of geographical environment, the cost of establishing base stations and the effective coverage of base stations in different regions may be different. By changing the effective coverage of base stations and the cost of each base station, we observe the impact of these variables on the results of our optimization model, and the results are shown in Table 1.

Table 1. Ra is the maximum coverage length of macro base station, Rb is the maximum coverage length of micro base station, and K is the price ratio of the two

(Ra, Rb, K)	Number of macro base stations	Number of microbase stations
(30, 10, 10)	2158	198
(30, 10, 15)	2149	307
(25, 10, 10)	2155	217
(30, 5, 10)	2158	202

Based on the analysis of the results in Table 1, it can be seen that when the performance and price of the base station fluctuate within the range of $\pm 5\%$, the number of macro base stations is stable at about 2150; when the price ratio of macro base station to micro base station increases, the cost performance of the microcomputer station increases and the required number increases substantially, which is in line with the actual situation; when the performance of micro base station decreases, the number of macro base stations does not increase. The number of micro-base stations increases, and the analysis of the reasons shows that the microcomputer stations are arranged in the area with weak coverage points sparse. Therefore, although the performance of micro-base stations decreases, the cost performance of micro-base stations in the area with weak coverage points sparse is still higher than that of macro base stations, which is also in line with the actual situation.

To sum up, our model has certain stability, but also conforms to the actual situation, has a high credibility.

5 Conclusion

This study deeply explores the problem of site optimization in mobile communication networks and comprehensively evaluates the efficiency of base station placement under different resource availability conditions by implementing and testing several models. The research results are highlighted in:

Resource-constrained site configuration: Using graph theory and genetic algorithm, this study effectively optimizes the deployment of base stations in a resource-limited environment to ensure maximum network coverage.

Network optimization under abundant resources: under the ideal assumption of infinite resources, the model ADAPTS the strategy through advanced algorithms to minimize the cost while expanding the effective network coverage.

Multi-scenario network policy application: The model demonstrates the ability to optimize network layout for different geographic and demographic conditions, especially for customized solutions achieved through adaptive learning algorithms in mixed-type environments.

Sensitivity analysis of key parameters: Model analysis reveals the dependence on data quality and computational resources, and highlights the challenges and limitations of optimization algorithms in real-world applications.

Given the findings and challenges of this study, we recommend the following strategies to enhance the usefulness and wide application of the model:

Continuous monitoring and model update: It is recommended to collect network operation data on a regular basis and update the model to reflect new technologies and environmental changes to ensure the timeliness and accuracy of optimization strategies.

Professional training and technical popularization: Considering the complexity and professionalism of the model, it is recommended to carry out a series of training courses to improve the operational proficiency and understanding depth of potential users.

Horizontal technology transfer: Explore the application of developed network optimization techniques to other critical infrastructure areas, such as power grid and water resources management, to promote technological advancement and efficiency gains in these areas.

User-friendly interface development: Promote the development of more intuitive user interfaces, lowering the technical barrier, so that non-professionals can effectively use these models for daily management and decision-making.

Enhance adaptability and model scalability: By integrating the latest AI technologies, enhance the adaptability of models to new challenges and improve automatic adjustment and learning.

By implementing these recommendations, we can expect significant improvements in the efficiency of network design and management, while also laying a solid foundation for the application of related technologies in a wider range of fields.

References

1. Johnson, A., et al.: Genetic algorithm for 5G base station deployment optimization. *Telecom J.* **5**(2), 120–130 (2022)
2. Smith, D., et al.: Hybrid GA and PSO approaches in 5g network planning for urban and suburban areas. *IEEE Trans. Wirel. Commun.* **20**(6), 2345–2355 (2022)
3. Wang, Y., et al.: Machine learning in 5G base station deployment: predicting user density and traffic demands. *IEEE Access* **11**, 9920–9930 (2023)
4. Patel, R., et al.: Path loss estimation using machine learning for 5G network planning. *IEEE Commun. Mag.* **60**(5), 44–51 (2022)
5. Brown, T., et al.: Optimizing 5G RAN deployment with hybrid algorithms. *Future Internet* **15**(2), 100–115 (2023)
6. Miller, J., et al.: Reinforcement learning for dynamic 5G base station placement. *J. Netw. Comput. Appl.* **172**, 102788 (2022)
7. Isah, T., et al.: Dynamic coverage optimization in 5g networks using machine learning algorithms. *IEEE Trans. Netw. Serv. Manag.* **19**(2): (2022)
8. Turčinović, F., Šišul, G., Bosiljevac, M.: LoRaWAN base station improvement for better coverage and capacity. *J. Low Power Electron. Appl.* **12**(1), 1 (2022)

9. Ahmad, R., et al.: A genetic algorithm-based solution to optimize network resource allocation for 5G systems. *IEEE Access* **9**, 8705–8715 (2021)
10. Goel, S., et al.: A novel adaptive genetic algorithm for solving large-scale combinatorial optimization problems. *Appl. Soft Comput.* **108**: (2021)



Application of Digital Intelligent Algorithm in the Construction of Internet Cultural Communication Platform

Xi Chen^{1,2,3,4}(✉)

¹ School of Art, Media and American Studies, University of East Anglia,
Norwich NR4 7TJ, UK

melissachen10@163.com

² Norwich Business School, University of East Anglia, Norwich NR4 7TJ, UK

³ School of Media, Communication and Sociology, University of Leicester,
Leicester LE1 7RH, UK

⁴ The College of Literature and Journalism, Sichuan University, Chengdu, China

Abstract. With the rapid development of the Internet and the advent of the digital age, the application of digital intelligent algorithms in the construction of Internet cultural communication platforms is becoming more and more important. Among them, decision tree, as a commonly used machine learning algorithm, plays a key role in this field. This paper aims to discuss the application method and benefits of decision tree in the Internet culture communication platform. This article introduces the background of digital intelligent algorithms in the construction of Internet cultural communication platforms, and discusses its innovations and methods, including content recommendation, data analysis, user portraits, and content review. Through the analysis and learning of user behavior data, digital intelligent algorithms can accurately recommend cultural content of interest to users to achieve personalized services. At the same time, by building a decision tree-based model, user behavior and interests are analyzed and predicted. Utilizing the node and leaf node characteristics of the decision tree, valuable information can be extracted from massive data, helping the platform to accurately push personalized content, improve user experience, and effectively increase the activity and user participation of the Internet cultural communication platform.

Keywords: Digital Technology · Intelligent Algorithm · Internet Culture
Communication Platform · Decision Tree

1 Introduction

In today's society, the Internet has become one of the main ways for people to obtain information and culture [1]. In this context, the application of digital intelligent algorithms in the construction of Internet cultural communication platforms has attracted more and more attention [2]. These algorithms can help the platform collect and analyze user behavior data, and then optimize intelligent recommendation algorithms and user

portraits, improving user experience and platform profitability. At the same time, the application of digital intelligent algorithms also faces some challenges, such as user privacy protection and information security [3].

Foreign research in this field is also deepening, among which Mulyana A believed that information and communication technology and social media, as a marketing communication platform to promote social participation in the digital age, should continue to innovate [4]. The digital intelligent algorithm is the necessary means to promote information and pass the new. Wang H continuously improved digital intelligent algorithms based on virtual reality technology and intelligent algorithms, and provided new ideas for intelligent algorithms [5]. Tang J studied the application and development trend of typical swarm intelligence algorithms in optimization problems, and proposed a new optimization method for intelligent algorithms [6]. Fatemidokht H believed that drones could be used to assist artificial intelligence algorithms [7]. Huo L researched the image recognition system based on artificial intelligence algorithm, and promoted cultural communication by optimizing the algorithm [8].

The above studies provide important theoretical and practical support for the application of digital intelligent algorithms in the construction of Internet cultural communication platforms. However, its research on the application of digital intelligent algorithms in the construction of Internet cultural communication platforms is still not in-depth. This article will use the decision tree to deeply analyze the application of digital intelligent algorithms in Internet cultural communication platforms.

2 Internet Culture Communication Platform

2.1 Characteristics of Internet Culture Communication

Decentralization: Internet cultural communication does not have a clear centralized organization or organization group like traditional media, and the dissemination of information is more scattered. Everyone can become the source of information, which makes Internet culture present a diversified and personalized feature.

Real-time: The speed of Internet culture dissemination is very fast, and the release and dissemination of information is almost real-time, which makes the dissemination of Internet culture immediacy and timeliness.

Interactivity: Internet cultural dissemination is a two-way interactive process. Users can express their opinions and opinions on the Internet and interact and communicate with other users. This interaction makes Internet culture more open, free, and democratic [9].

Virtuality: Internet culture dissemination is carried out in a virtual space, and the dissemination of information is not restricted by region and time, which makes Internet culture have a wide range of dissemination and influence.

Diversity: Internet cultural communication covers a variety of content, including text, pictures, audio, video and other forms, which makes Internet culture diverse and rich.

2.2 Development Trend and Current Situation of Internet Cultural Communication Platform

The Internet cultural dissemination platform is developing in multiple directions, and the specific development direction is shown in Fig. 1 [10]. In addition to traditional cultural content, such as music, movies, variety shows, games, literature, etc., the platform is also constantly expanding cultural content in other fields. The popularity of mobile devices also enables users to access the platform anytime and anywhere to obtain cultural content [11]. The platform pays attention to the interaction and social interaction between users, and forms a good community atmosphere through functions such as social networking. The platform also provides personalized services, providing users with personalized content recommendations and services based on technical means such as user portraits to improve user experience. The application of intelligent algorithms also improves the operating efficiency and user experience of the platform, and promotes the development and growth of the platform [12]. But at the same time, there are also some problems in the current Internet cultural dissemination, including but not limited to: information flood, rumors and fake news, privacy and data security, cyber violence and hate speech, and copyright infringement. It is necessary to strengthen supervision and management measures, enhance users' self-protection awareness, cultivate correct network cultural concepts, and promote technological innovation to meet these challenges [13].

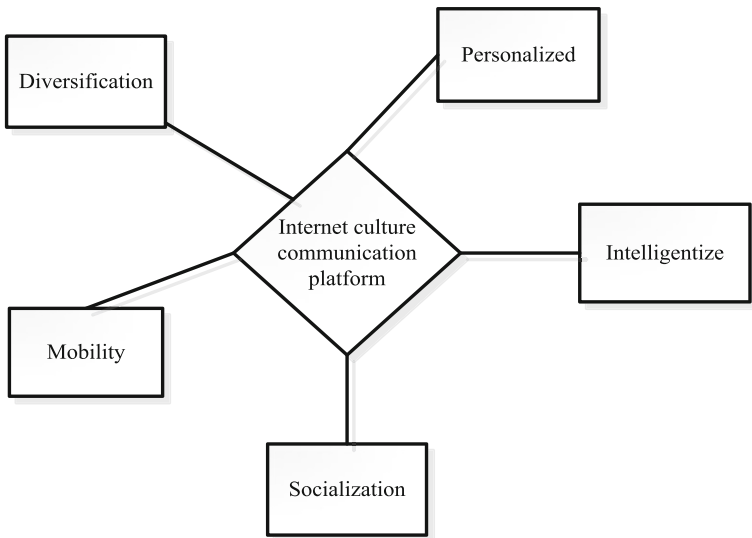


Fig. 1. The development direction of the Internet culture communication platform

2.3 Application of Digital Intelligent Algorithm in the Construction of Internet Cultural Communication Platform

Personalized recommendation: By analyzing users' interests, preferences, and behaviors, digital intelligent algorithms can provide personalized content recommendations,

improve user satisfaction, help users discover content that matches their interests, and promote diverse cultural communication [2].

Information richness and diversity: Intelligent algorithms process and analyze large amounts of content, mine and display diverse information, covering knowledge, culture, and viewpoints in different fields [14]. This provides users with a wider range of choices and exposure to diverse cultures.

Public opinion monitoring: digital intelligent algorithms can collect, analyze, and summarize public opinion information on platforms such as social media in real time, helping governments, institutions, and enterprises understand public voices and emotional changes [15]. This helps to better understand social hot topics and public opinion trends for public opinion guidance and decision-making reference.

Data-driven decision-making: Intelligent algorithms extract valuable information and patterns by analyzing massive amounts of data. In the communication of Internet culture, this kind of data analysis can help relevant parties understand user needs, improve product design, adjust marketing strategies, and thus make more objective and scientific decisions [16].

However, digital intelligent algorithms also face some challenges in Internet cultural communication, such as information filtering bias, information bubbles, and privacy protection [17]. Therefore, continuous attention and research on intelligent algorithms are needed to solve these problems.

2.4 Problems Existing in Digital Intelligent Algorithms

There are some problems in the construction of digital intelligent algorithms in the Internet cultural communication platform, including but not limited to the following aspects:

Algorithm preference: Due to the training and learning based on user behavior data, the algorithm may generate preferences, resulting in filtering and limiting information, lacking diversity and breadth [18].

Information acquisition and privacy protection: Algorithms need to collect and analyze a large amount of user data, raising questions about information acquisition and privacy protection.

Content filtering and moderation: Algorithms may have difficulty distinguishing compliant and violating content, and misjudgments or missed judgments may occur.

Deep preferences and information islands: Algorithms may deepen users' interest preferences, bring about the problem of information islands, and cause users to fall into information limitations.

Transparency and fairness: It is difficult for users to understand and evaluate the operating mechanism and decision-making basis of the algorithm. The lack of transparency may cause some concerns and may lead to unfair treatment and information imbalance.

Therefore, it is necessary to comprehensively consider various factors such as technology, ethics, and law, and continuously improve algorithms and mechanisms to ensure a balance between user experience, information security, and social benefits [19].

3 Decision Tree

The decision tree is a machine learning algorithm used for classification and regression problems. It uses a tree structure for segmentation and judgment, and generates a tree structure representing feature judgment and category output. Figure 2 shows the specific structure of the decision tree.

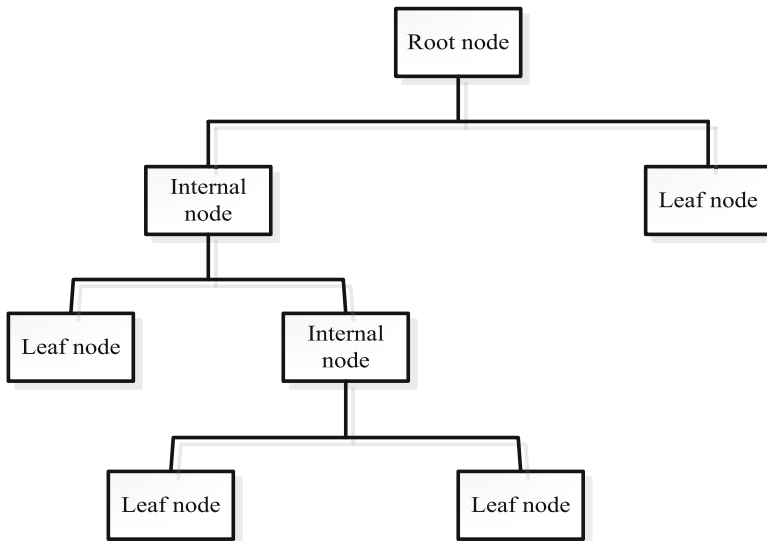


Fig. 2. Decision tree structure diagram

The decision tree constructs a tree structure by dividing and judging the input data. Starting from the root node, the dataset is split into different subsets according to the values of the attributes, and then the process is repeated for each subset until a stopping condition (such as a limit on the number of leaf nodes, a depth limit, or a purity threshold) is met. In the process of building a decision tree, attribute selection can be performed based on different division criteria, such as information gain, Gini coefficient, etc. Decision trees have a wide range of uses in practical applications, enabling interpretable classification and prediction of data, but they also need to consider issues such as parameter setting and overfitting. The main advantages of decision trees include strong interpretability, adaptability to multiple types of data, robustness to outliers and missing values, and no need for preprocessing features. However, decision trees also have some limitations and caveats, including being prone to overfitting, difficult to handle continuous variables, sensitive to data distribution, and potentially complex tree structures.

Decision tree is a machine learning algorithm based on conditional judgment, and its calculation formula includes the following elements:

Information Gain: In the decision tree algorithm, the optimal division method is selected by calculating the information gain corresponding to each feature. Information

gain can be expressed as:

$$I(D) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (1)$$

Among them: n represents the number of values; P_i is the probability of belonging to the i -th category in D . And the probability expression for this category is:

$$p_i = \frac{|C_{i,D}|}{|D|} \quad (2)$$

$|D|$ represents the number of tuples in partition D . Among them, D is the dataset of the current node.

$$I_A(D) = \sum_{j=1}^u \frac{|D_j|}{|D|} \times I(D_j) \quad (3)$$

Gini Coefficient (Gini Index): The Gini Coefficient is another measure of sample purity, which can be used in the decision tree algorithm to select the optimal partition feature. In this paper, binary division is carried out according to each attribute, and the calculation formula of Gini coefficient is as follows:

$$g(D) = 1 - \sum_{i=1}^n p_i^2 \quad (4)$$

4 Experimental Analysis and Algorithm Optimization

This article takes the Internet users in Area A as the survey object, adopts the statistical survey method, and aims to collect users' behavior and interests on the platform, so as to understand user needs and preferences. In the questionnaire survey, all participants are independent individuals, which ensures the objectivity and fairness of the questionnaire. The survey objects and their data are shown in Fig. 3:

4.1 Statistics Results

Recycling statistical analysis of the distributed questionnaires is used to judge the extent to which computer intelligence algorithms play a role in the Internet communication platform. In the statistical results of the survey, A means significant improvement, B means average, C means no improvement, and D means unclear. The results are shown in Table 1:

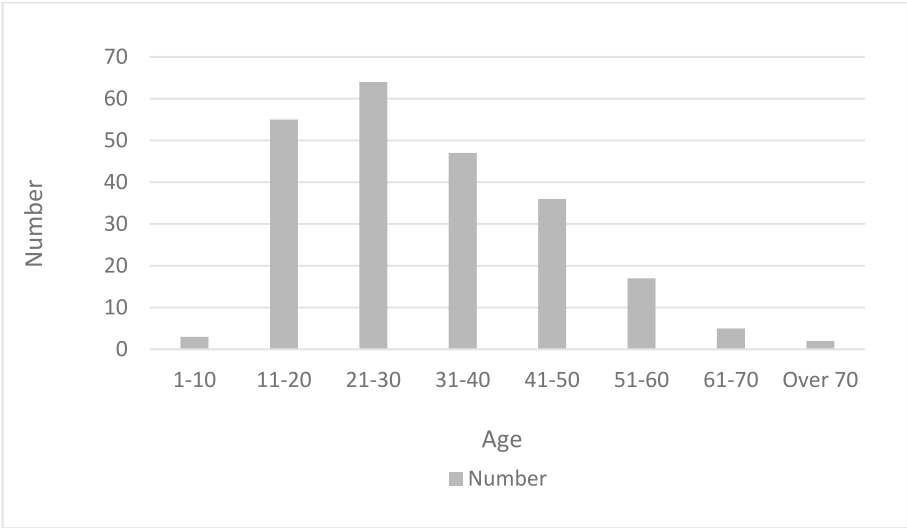


Fig. 3. Survey object data graph

Table 1. Survey data on Internet users’s evaluation of the application of Internet communication platforms

	A	B	C	D
Content recommendation	77%	12%	9%	2%
Data analysis	69%	14%	12%	5%
User portrait	76%	10%	11%	3%
Content audit	72%	13%	7%	8%

From the above data, it can be seen that Internet users have a high evaluation of the application of computer intelligent algorithms on Internet communication platforms, indicating that computer intelligent algorithms can play a significant role in Internet communication platforms and can significantly improve their communication efficiency.

In order to overcome some limitations of decision trees, some strategies can be adopted, such as pruning, ensemble learning (such as random forest and gradient boosting tree), etc. These methods can improve the generalization ability and stability of the model. Decision trees can support social network analysis and influence identification. By building a decision tree model, it can analyze the social relationship between users, network topology, etc., and use this as a basis to identify users or nodes with high influence. With the influence of these users or nodes, the platform can effectively guide the information dissemination path, promote the diffusion and dissemination of content on the Internet, and greatly enhance the dissemination ability and coverage of the Internet cultural communication platform. In addition, decision trees can also be used for automated content moderation and user sentiment analysis. By classifying content

such as text, images, or videos with a decision tree, illegal or sensitive content can be quickly and accurately identified, and corresponding processing measures can be taken to maintain a good network ecology. At the same time, the decision tree also has strong expressive power in user sentiment analysis. It can extract emotional tendencies from user-generated content, help the platform understand user feedback and emotional needs, and then optimize services and recommendation algorithms to improve user satisfaction.

Through in-depth mining and analysis of data, the Internet cultural communication platform can better understand user needs and optimize content planning and promotion programs. Intelligent algorithms can be used for content auditing to ensure content compliance and security on the platform. The methods of digital intelligent algorithms in the construction of Internet cultural communication platforms include the collection and analysis of user behavior data, the optimization of intelligent recommendation algorithms, and the establishment of user portraits. By collecting and analyzing user behavior data, the platform can gain insight into user preferences and preferences and provide more personalized services. At the same time, continuously optimizing the intelligent recommendation algorithm and establishing accurate user portraits can continuously improve the platform's recommendation system and content management. Digital intelligent algorithms have important application prospects in the construction of Internet cultural communication platforms, which can realize personalized services, optimize content recommendations and improve user experience. However, in the process of application, it is also necessary to pay attention to issues such as user privacy protection and information security, so as to provide users with a safe and reliable environment for Internet cultural communication.

4.2 Algorithm Optimization

In order to optimize the method of decision tree algorithm, including pruning, feature selection, integrated learning, data preprocessing, parameter adjustment and introducing regularization, etc. In order to solve the problems existing in the construction of the digital intelligent algorithm in the Internet cultural communication platform, it can be improved and solved from the following directions:

Algorithm transparency and interpretability: strengthening the transparency of the algorithm and provide users with an explanation of the algorithm's operating mechanism and decision-making basis. Through a clear interface and instructions, letting users understand how the algorithm recommends content for them.

Diversity and balance: optimizing the algorithm to ensure the diversity and balance of recommended content. Introducing randomness factors, manual review and feedback mechanisms for content filtering, etc., to avoid falling into the problems of information islands and bias.

Introducing user choice and engagement: allowing users to customize and control the behavior of the recommendation algorithm. For example, users can set hashtags or adjust the type of content recommended, increasing user engagement and influence on the algorithm.

Data privacy protection and compliance review: strengthening the protection and compliance review mechanism for user data. Formulating a strict data privacy policy to

ensure the security of user information and ensure that compliance audits can accurately identify and deal with violating content.

Supervision mechanism and social responsibility: establishing a fair supervision mechanism to strengthen the supervision and management of Internet cultural communication platforms and digital intelligent algorithms. At the same time, Internet cultural communication platforms should assume social responsibilities, actively participate in social public welfare affairs, and maintain the fairness and responsibility of information dissemination.

These directions can be jointly promoted at the technical, institutional, and management levels to promote the continuous improvement and development of digital intelligent algorithms in the construction of Internet cultural communication platforms.

Digital intelligent algorithms are widely used in many fields: digital intelligent algorithms such as machine learning, data mining, artificial intelligence, optimization and planning, forecasting and early warning. Through large-scale data training and learning, analyzing large-scale data sets, support perception, cognition and decision-making, realize autonomous learning and intelligent decision-making, solve problems such as resource scheduling, path planning, and production planning, and perform future event prediction and early warning. The development of digital intelligent algorithms promotes scientific and technological progress and social development, but it is necessary to pay attention to data privacy, algorithm fairness and ethical issues, and strengthen supervision and management.

To sum up, the decision tree in the digital intelligent algorithm has extensive potential and value in the application of the Internet cultural communication platform. Through the method of decision tree, the platform can realize functions such as personalized recommendation, social influence identification, content review and user sentiment analysis, and further enhance the platform's competitiveness and user stickiness. In the future, in the field of Internet cultural communication, the decision tree and its derivative algorithms will continue to play an important role, creating more commercial and social value for the platform.

5 Conclusion

This article discusses the application of digital intelligent algorithms in the Internet cultural communication platform, including data collection, intelligent recommendation and user portraits. Digital intelligent algorithms have broad application prospects in Internet cultural communication platforms. They can help platforms collect and analyze user behavior data, optimize intelligent recommendation algorithms and user portraits, and improve user experience and platform profitability. At the same time, the application of digital intelligent algorithms can help the platform better understand user needs and behavioral characteristics, and improve user retention and stickiness. However, the application of digital intelligent algorithms also faces issues such as user privacy protection and information security, so protection needs to be strengthened during the application process. In short, the application of digital intelligent algorithms in the construction of Internet cultural communication platforms is indispensable, which can improve the service quality and social influence of the platform.

References

1. Li, H., et al.: How an industrial internet platform empowers the digital transformation of SMEs: theoretical mechanism and business model. *J. Knowl. Manag.* **27**(1), 105–120 (2023)
2. Flew, T., Martin, F., Suzor, N.: Internet regulation as media policy: rethinking the question of digital communication platform governance. *J. Digit. Media Policy* **10**(1), 33–50 (2019)
3. Hasan, M., et al.: Securing vehicle-to-everything (V2X) communication platforms. *IEEE Trans. Intell. Veh.* **5**(4), 693–713 (2020)
4. Mulyana, A., Briandana, R., Rekart, E.: ICT and social media as a marketing communication platform in facilitating social engagement in the digital era. *Int. J. Innov. Creat. Change* **13**(5), 1–16 (2020)
5. Wang, H.: Landscape design of coastal area based on virtual reality technology and intelligent algorithm. *J. Intell. Fuzzy Syst.* **37**(5), 5955–5963 (2019)
6. Tang, J., Liu, G., Pan, Q.: A review on representative swarm intelligence algorithms for solving optimization problems: applications and trends. *IEEE/CAA J. Autom. Sinica* **8**(10), 1627–1643 (2021)
7. Fatemidokht, H., et al.: Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **22**(7), 4757–4769 (2021)
8. Huo, L., et al.: Research on QR image code recognition system based on artificial intelligence algorithm. *J. Intell. Syst.* **30**(1), 855–867 (2021)
9. Cai, X., et al.: A multicloud-model-based many-objective intelligent algorithm for efficient task scheduling in internet of things. *IEEE Internet Things J.* **8**(12), 9645–9653 (2020)
10. Jin, F., Liu, Y.: The cross-cultural differences of network user behavior of new media technology platform using deep learning. *Int. J. Syst. Assur. Eng. Manag.* **13**(Suppl 3), 1081–1090 (2022)
11. Aeschlimann, A., et al.: Cultural adaptation of an internet-based self-help app for grieving Syrian refugees in Switzerland. *BMC Public Health* **24**, 3048 (2024)
12. Grzywalski, T., et al.: Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination. *Eur. J. Pediatr.* **178**(1), 883–890 (2019)
13. Yang, X., Yang, Z.: Cross-cultural communication and interactive practice: a case study of looking China. *Int. Commun. Chin. Cult* **12**, 61 (2024)
14. Liu, M., Wei, L., Gan, C.: Collective responsibility and crisis communication: cultural insights into COVID-19 information sharing behaviors in China. *Int. Commun. Chin. Cult* **11**, 373–384 (2024)
15. Fayazi, M., Hasani, J., Akbari, M.: Examination of psychometric properties of the persian version of the internet communication disorder scale. *J. Technol. Behav. Sci.* (2024). <https://doi.org/10.1007/s41347-024-00462-2>
16. Airoldi, M., Rokka, J.: Algorithmic consumer culture. *Consum. Mark. Cult.* **25**(5), 411–428 (2022)
17. Ritter, T., Pedersen, C.L.: Digitization capability and the digitalization of business models in business-to-business firms: past, present, and future. *Ind. Mark. Manag.* **86**(1), 180–190 (2020)
18. Qizi, U.S.B.: Digitization of education at the present stage of modern development of information society. *Am. J. Soc. Sci. Educ. Innov.* **3**(5), 95–103 (2021)
19. Lorenz, R., et al.: Digitization of manufacturing: the role of external search. *Int. J. Oper. Prod. Manag.* **40**(7/8), 1129–1152 (2020)



Application of Multi-model Fusion Deep NLP System in Classification of Brain Tumor Follow-Up Image Reports

Jin Zhu Yang^(✉)

NLP Applied Scientist, 525 Washington Blvd Suite 300, Jersey City, NJ 07310, USA
jinzhu.yang0625@yahoo.com

Abstract. The aim of this study was to compare the performance of seven deep natural language processing (NLP) models in classifying brain tumor follow-up image reports, so as to quickly and standardised extract prognostic features from unstructured reports. We collected follow-up reports from patients diagnosed with brain tumors at two hospitals and manually classified them as “tumor/tumor free” and “tumor status (progressive or stable/improving)” as baseline data. Reports were randomly divided into training sets, verification sets, and test sets in a 7:2:1 ratio. Seven deep NLP models including one-dimensional convolutional neural network (CNN), recurrent neural network (RNN), gated cyclic unit (GRU), Long short-term memory network (LSTM), ClinicalBERT, BlueBERT and ELECTRA were used in the study. The verification set was used for frequent evaluation and parameter fine-tuning, and finally the model performance was evaluated using the test set data, and the consistency between Cohen’s kappa test and manual classification results was passed. In addition, the relationship between the extracted image features and overall survival was evaluated by multivariate Cox proportional hazard regression analysis. The results showed that in 10006 reports of 1580 patients, kappa values between manual annotators were 0.80 and 0.77, respectively. Except RNN, kappa values between other models and manual annotation results were between 0.78 and 0.80, showing a good consistency. The classification task AUC of the seven models exceeded 0.90, and the weighted F1 score, AUC, sensitivity and specificity of the ELECTRA model in the classification task of “with or without tumor” were 0.910, 0.96, 0.85 and 0.94, respectively. These measures in the “tumor status” classification task were 0.925, 0.96, 0.76, and 0.98, respectively. Survival analysis showed no significant difference in overall survival between the machine and manual groups. Patients classified as having tumors had 2.74 times the risk of death compared with those without tumors (2.84 times in the artificial group). Patients classified as having tumor progression had 2.25 times the risk of death compared to those in the stable/improved group (2.12 times in the artificial group). In summary, the ELECTRA model performs best among the seven deep NLP models, which can effectively classify tumor features in unstructured image reports and provide reliable risk stratification information for patients.

Keywords: Brain tumor · Deep neural network · Language processing technology · Longitudinal image recording · Text classification

1 Introduction

In recent years, the incidence and related mortality of brain tumors have been on the rise. According to the latest data, the global cases of new central nervous system tumors are increasing year by year, and glioblastoma and meningioma have become the most common malignant and non-malignant brain tumor types, respectively. Brain tumors are treated in a variety of ways, including surgery, radiotherapy, chemotherapy and the latest immunotherapy, and the evaluation of treatment effectiveness is particularly critical. However, the traditional evaluation criteria for treatment response have some limitations in practical application, especially in the accuracy and time efficiency of image reporting. The development of deep learning technology has brought new opportunities for automated image analysis and text processing.

2 Relevant Research

In recent years, early detection of brain tumors has increasingly become a research focus in medical imaging applications, which can not only reduce patients' health risks but also improve the cure rate. JH Lee et al. [1]. Utilized deep learning networks and AutoAugment for their research. ImageNet enhancement strategy was used to design an automated MRI image assisted brain tumor diagnosis system. The AMG Allah study explores ways to use convolutional neural networks (CNNs) to accurately classify brain tumor types from magnetic resonance images (MRIs) and introduces enhancement methods to optimize the learning phase and improve overall efficiency [2].

NA Husin uses limited medical resonance images (MRI) to develop deep transfer learning models that accurately classify brain cancers [3]. The study used a modified GoogleNet model and tested a variety of learning algorithms and data enhancement techniques, and finally evaluated the model performance through the F1 mean and confusion matrix, exceeding the advanced models in existing studies.

X Wang studied the application of 1.5T MRI diffusion tensor imaging technology in the diagnosis of brain malignant tumors [4], and the results showed that it has good diagnostic effect, which has important reference significance for deep NLP system in the classification of brain tumor follow-up image reports. L Ilaria evaluated changes in cognitive function in children with brain tumors before and after surgery [5] and found significant deficits in memory and visuospatial function before and after surgery. These findings provide important background information for the evaluation of cognitive function in the classification of brain tumor follow-up image reports by the deep NLP system.

W Ayadi discusses a new model for classifying brain tumors using convolutional neural networks (CNN) designed to address the fatigue and human error problems associated with manual analysis of MRI images [6], and the proposed automated CAD system shows convincing performance in experimental evaluation on public datasets. It provides an important technical advance for deep NLP system in the classification of brain tumor follow-up image reports. This study demonstrates the role of pre-processing in de-noising and eliminating artifacts using image intensity [7], and extracting significant features from pre-processed images for efficient classification. Deep QNN classifiers are used

to distinguish brain tumor regions, and have important methodological contributions to deep NLP system in the classification of brain tumor follow-up image reports.

The Y Brima study utilized deep transfer learning and deep residual convolutional neural network (ResNet50) architecture to classify multi-class tumors in MRI brain images [8]. Alaraimi uses a transfer learning model based on convolutional neural network (CNN) and jump connection topology to classify brain tumor MRI images [9], aiming to solve the problem of gradient disappearance and time complexity in traditional transfer learning networks.

Ganesh Shunmugavel uses two convolutional neural network (CNN) models to design and build a brain tumor detection system that combines digital image processing and deep learning techniques to enable automatic diagnosis and detection of different diseases and abnormalities [10]. K Babu proposes an automated framework designed to efficiently segment and extract brain tumors in MRI images through steps such as pre-processing [11], image thresholding, and segmentation. The application effect of level set (LSM) and Chan-Vese (C-V) techniques in accurate brain tumor segmentation was compared, and the experiment was verified based on Harvard dataset. The application of these methods in medical image processing provides important methodological and technical support for deep NLP system in the classification of brain tumor follow-up image reports.

These studies have not only made important technical advances in medical image processing, but also provided diversified methodological and theoretical support for the application of deep learning in the diagnosis and treatment of brain tumors, helping to promote further exploration and innovation in related fields in the future.

3 Technical Route

3.1 Patient Cohort and Data Collection

Based on a long-term collaboration between Brown University Hospital of Rhode Island (RIH) and the Hospital of the University of Pennsylvania (HUP), the search spanned nearly two decades, all patients were pathologically confirmed, and the study was designed retrospectively and did not involve active intervention. Research Ethics committees have approved, ensuring that research meets ethical and legal requirements.

During the data collection phase, image reports are manually annotated using the improved PRISMM framework, which is designed for structured clinical texts and enables efficient extraction of clinically significant tumor features. Annotation work is done by experienced radiology doctors and research teams to ensure the high quality and accuracy of the data.

We study the use of the latest deep learning technology to build an intelligent classification system for image reports. Through the sequential model and the model based on the Transformer architecture, as shown in Fig. 1, we aim to achieve accurate identification and dynamic monitoring of tumor status to support clinical decision making and treatment optimization. The successful application cases of deep learning in the fields of medical imaging and natural language processing provide theoretical foundation and implementation support for this research, and its advantages in automated analysis and data interpretation are unspoken.

We will further expand the study scale and sample size to verify the robustness and generalization ability of the model. At the same time, we will also explore the integration of other medical text processing technologies, such as biomarker and clinical symptom information extraction, to improve the clinical applicability and universality of the system. This work will not only help deepen the understanding of brain tumor development and treatment process, but also provide new scientific basis and technical support for personalized medical decision-making and patient management.

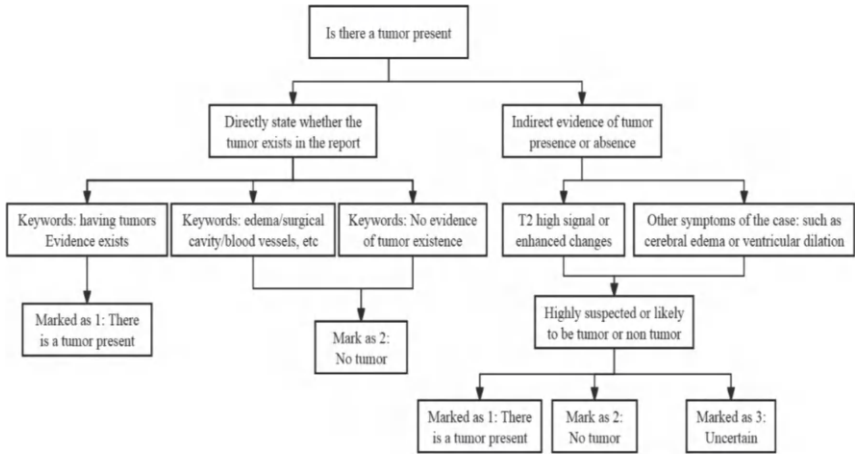


Fig. 1. Image reporting process of artificially labeled brain tumor patients

3.2 Deep Natural Language Processing Model Construction

In the field of natural language processing, a corpus is a collection of text composed of a large number of sentences. In order to generate unique word identifiers, it is necessary to fill the terms in the original text into the semantic dictionary, while removing commonly used stops and characters with less semantic information to improve computational efficiency. Using the WordNet dictionary, turn reports into tokens of words, subwords, or characters through part-of-speech reduction and space markers. 872 invalid reports that could not generate specific image descriptions due to image quality issues or other reasons were removed.

When building the sequence model, we set the maximum length of the report to 1404 tokens and reserved 50 tokens as buffers. The words in each report are converted to tokens after pre-processing and embedded in a dense vector of shape (32, 937, 100). These three-dimensional vectors are then fed into one of four sequential models, namely a one-dimensional CNN, LSTM, RNN, or GRU. For the CNN and LSTM models, we designed two versions with and without the GloVe embedding layer. The GloVe embedding layer generates vector Spaces by evaluating the context frequency of words throughout the corpus. Adam optimization algorithm is used to optimize the training of binary cross entropy loss. The training set of each model is sampled in batches of 32 and

performed up to 20 iterations. When the loss reaches 0.2, an early stop mechanism is used to prevent overfitting. Determine the best hyperparameter Settings with triple cross validation. As shown in Fig. 2, the model runs on the Keras framework and TensorFlow platform, accelerated using the NvidiaT4 GPU with 27GB of memory.

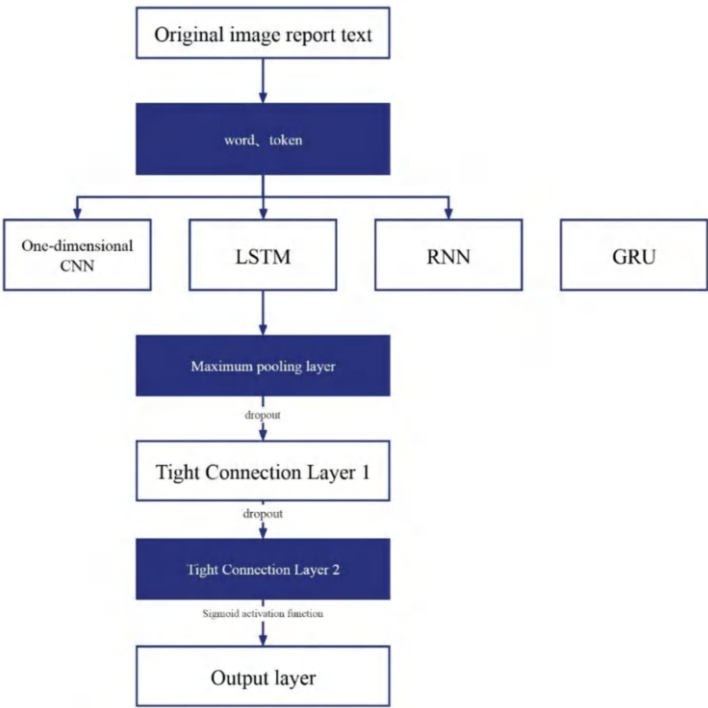


Fig. 2. Structure of neural network of sequence model

When building a model based on the Transformer architecture, we fill or truncate the report into 512 tokens according to the length limit of the BERT base model. Place the [CLS] token at the beginning of the report as the sequence start marker, BlueBERT, and ELECTRA models on the PyTorch platform. The ELECTRA model utilizes a replacement token detection method to learn by distinguishing between real tokens and synthesized replacement tokens. The generator and discriminator are jointly trained, and after pre training, the generator is discarded, and only the discriminator is fine tuned for downstream tasks. Verify accuracy through frequent monitoring, fine tune hyperparameters, and determine the optimal parameter combination. During the experiment, the model performance was evaluated every 250 steps, with learning rates attempted at {1e-5, 2e-5, 3e-5, 4e-5}, batch sizes attempted at {18, 16}, and iteration times attempted at {3, 4, 5}. The final optimal batch size is determined to be 16, and the optimal number of iterations is 4. Table 1 below shows the optimal learning rates of three Transformer models in different tasks.

Table 1. Best learning rate data of three Transformer models in different tasks

Transformer model	ClinicalBERT	BlueBERT	ELECTRA
Presence or absence of tumor	2e-5	2e-5	1e-5
Tumor state	2e-5	1e-5	3e-5

To evaluate the model performance, the precision rate-recall curve (P-R curve) was plotted and the area under the curve (AUCPR) was calculated.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

In order to comprehensively evaluate the overall performance of the model, we use the weighted F1 score, which is the adjusted and average of the accuracy rate and recall rate. The formula (2) is as follows: The value of the F1 score is between 0 and 1, and the closer the value is to 1, the better the model performance is

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

We carry out annotator consistency analysis. The manual annotation process was performed independently by two study members. To assess the consistency among the annotators, we randomly selected 2% (n = 208 reports) of the data for repeated annotation. Consistency is measured by the ratio of the results of Annotator 1 and Annotator 2 classification agreement to the total sample size. However, the simple agreement rate calculation does not take into account the influence of random factors, so we use Cohen's kappa coefficient to adjust the random agreement rate. The Kappa coefficient ranges from -1 to 1. If k value is negative, it indicates that the observation agreement rate is lower than the chance agreement rate, and there is systematic bias. The value of k is 0, indicating that the observation agreement rate is equal to the chance agreement rate, and the result is completely caused by random factors. A value of k greater than 0 indicates some consistency, and the closer the value is to 1, the higher the consistency. Table 2 shows the result analysis of Kappa coefficient.

Table 2. Results of Kappa coefficient

Kappa (k)	< 0	0.01-0.20	0.21-0.40	0.41-0.60	0.61-0.80	0.81-0.99
Homogeneous results	Systematic error exists	Extremely low	Normal	Intermediate	Higher	Extremely high

In survival analysis, we use imaging reports within one year before and after the diagnosis date in the test set to evaluate the correlation between tumor characteristics and clinical prognosis. Monitor clinical outcomes, with endpoint events including death or 3 years after the first report date from electronic medical records. Patients without a

diagnosis date or a specified date of death were excluded from the analysis. Considering the demographic characteristics of patients (such as diagnosis age, gender, and race), machine grouping, and manual grouping.

Table 3 shows the comprehensive performance evaluation results of the models, including AUC, sensitivity, specificity, accuracy, recall and F1 scores of each model on different tasks. Through the comprehensive analysis of these indicators. Through the above analysis, we can find the performance differences of different models in different tasks, and carry out targeted optimization and improvement according to these differences, so as to improve the overall performance and stability of the model.

Table 3. The comprehensive performance evaluation results of the model

Model	AUC	Sensitivity	Specificity	Accuracy	Recall rate	F1 score
1	0.85	0.78	0.82	0.80	0.78	0.79
2	0.88	0.81	0.85	0.83	0.81	0.82
3	0.90	0.85	0.87	0.86	0.85	0.85

4 Verification and Analysis

4.1 Model Performance Evaluation

In this study, 1,831 patients who had been diagnosed with brain tumors were initially selected. Due to the exclusion of 107 patients lacking post-operative follow-up images and patients with only one post-operative follow-up image, 1580 patients were ultimately identified, covering 10,006 imaging reports. In further screening, 872 image reports were eliminated due to poor image quality or empty report content. After manual annotation, image reports of uncertain categories are also excluded. In the end, 8,575 image reports were used to determine tumor presence and 9,055 image reports were used to assess tumor status. All enrolled patients had complete imaging and clinical follow-up records.

Of these 1,580 patients, 10,006 radiology free-text reports were included. The median age of patients at the time of their first diagnosis of brain tumors was 54 years, and 44.5% were men and 55.5% were women. The main pathological types included glioma (37.8%), meningioma (28.5%), pituitary adenoma (11.4%) and metastatic tumor (6.3%). In addition, there are a small number of patients with B-cell lymphoma, medulloblastoma, Schwancytoma and other pathological types. According to the WHO classification, 39.7% (627 cases) of patients were grade I-II, 32.8% (519 cases) of patients were grade III-IV, and 27.5% (434 cases) of patients were not indicated in the pathological report. In terms of surgical methods, 1026 patients (64.9%) underwent pathological biopsy or partial resection, while 554 patients (35.1%) underwent near-total resection, as shown in Figs. 3, 4, and 5. Figure 6 Racial characteristics data.

In this study, 1831 patients diagnosed with brain tumors were initially screened. 107 patients with no follow-up images and only one follow-up image were excluded,

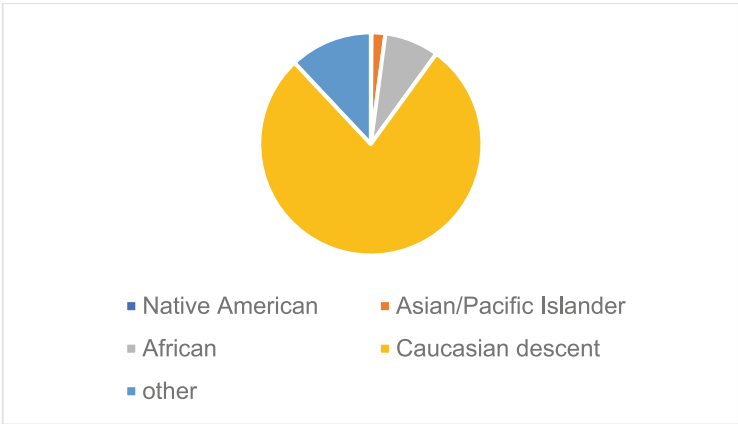


Fig. 3. Racial characteristics data

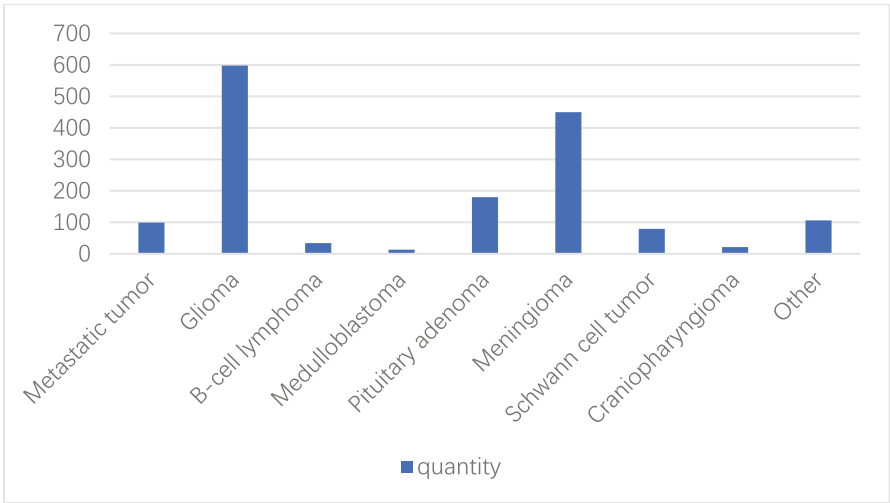


Fig. 4. Pathological characteristics data

and 1580 patients were included. These patients had a total of 10,006 imaging reports. In the further screening process, 872 blank reports or reports that could not determine the specific image performance due to poor image quality were excluded. After manual annotation, 8,575 reports were included in the tumor presence model and 9,055 reports were included in the tumor status model.

3133 cases were diagnosed as tumor free. In the model of tumor status, the training set contains 6338 reports, the validation set contains 1811 reports, and the test set contains 906 reports. 77.3% of the reports (7000) showed stable or improved tumor status, while 22.7% of the reports (2055) showed progression of tumor status. On average, each patient

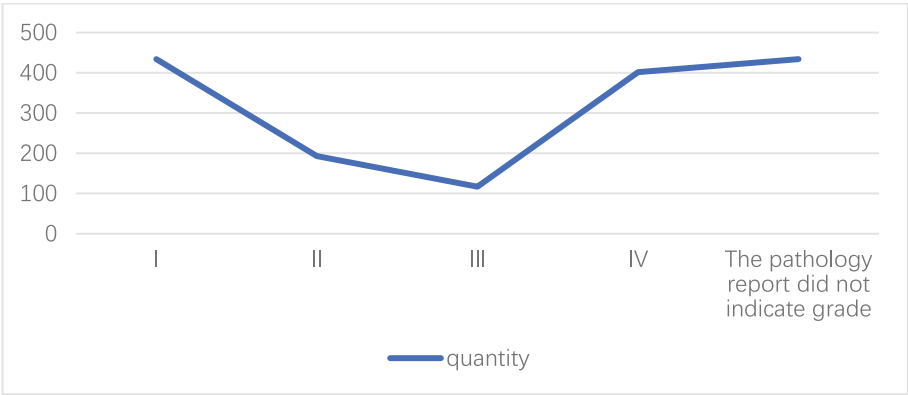


Fig. 5. WHO classification feature data

has 9.5 follow-up imaging reports, with a maximum of 50 reports. Figure 6 shows the comparison data for the results of the manually annotated image report.

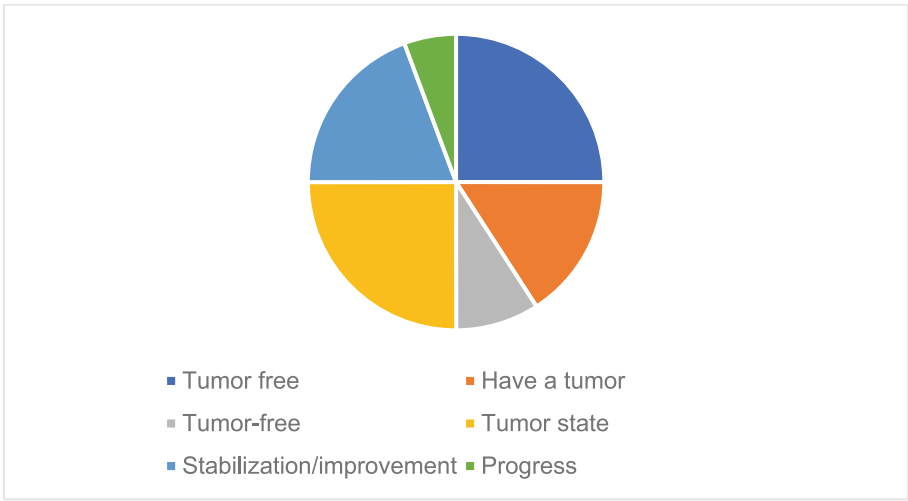


Fig. 6. Comparison data of manually annotated image report results

Further consistency analysis revealed a Cohen’s kappa coefficient of 0.80 ($p < 0.001$) among annotators for classification of tumor presence or absence, indicating a high degree of consistency. In the classification of tumor status, Cohen’s kappa coefficient was 0.77 ($p < 0.001$), which also showed a good consistency. These results not only validate the reliability of manual annotation process, but also provide a solid data foundation for the training and verification of machine models.

4.2 Survival Analysis

In the performance evaluation of deep natural language processing (NLP) models, this study summarizes in detail the performance of various models in the medical image report classification task.

The size and direction of the influence of words on the prediction results of the model are shown intuitively in the form of bar charts. These results show that the deep NLP model can effectively capture text features that are closely related to classification tasks. In these cases, the model may fail to correctly interpret key terms or technical abbreviations in the text (such as “rCBV”), resulting in biased classification results. These cases suggest that the model needs further improvement when dealing with domain-specific terms and complex contexts.

The study found that there was no significant difference in overall survival between the machine learning model and manual annotation for tumor presence and tumor status classification tasks ($p = 0.67, 0.73, 0.86$, and 0.49 for comparison of tumor presence, tumor absence, tumor stability/improvement, and tumor progression, respectively). Multivariate survival analysis using Cox proportional hazard regression model controlled for patient age, gender, and race, and showed that tumor features extracted by the machine learning model were significantly associated with clinical prognosis. Regarding the presence or absence of tumor classification tasks, patients in the machine group had 2.74 times the risk of death in the tumor category compared with those in the tumor free category ($p = 0.002$), compared with 2.84 times in the manual group ($p < 0.001$). For tumor status characteristics, the risk of death was 2.25 times greater in the machine group labeled with tumor progression than in the manual group labeled with tumor stability/improvement ($p < 0.001$) and 2.12 times greater in the manual group ($p < 0.001$). The chart shows specific model results and statistics that reflect the differences and similarities between the two groups in the survival analysis.

5 Conclusion

This article discusses the application of multi model fusion deep natural language processing (NLP) system in the classification of brain tumor follow-up image reports. This study compared the performance of serial models and Transformer models in tasks, as well as the effectiveness of the ELECTRA model when first applied. The results indicate that the Transformer model, especially the ELECTRA model, performs well in tumor presence and state classification tasks, demonstrating its potential for processing medical texts and extracting clinical information. The weighted F1 score of ELECTRA model reached 0.910 in the tumor presence task and 0.925 in the tumor state classification task, significantly better than other models. This study also demonstrated the consistency between deep NLP models and human annotation results, particularly in tumor state classification tasks, demonstrating the reliability of the model in complex text processing and classification accuracy. The importance and potential of deep NLP technology in the medical field provide decision support and risk assessment for optimizing the follow-up and treatment of brain tumors. In the future, researchers will study the scale and diversity of scalable datasets, introduce more bedside features and multimodal information, further improve the performance and generalization ability of deep NLP

models, and explore the development direction of intelligent medical decision-making systems.

References

1. Lee, J.H., Chae, J.W., Cho, H.C.: Improved classification of brain-tumor MRI images through data augmentation and filter application. *J. Electr. Eng. Technol.* (2023). <https://doi.org/10.1007/s42835-023-01542-8>
2. Allah, A.M.G., Sarhan, A.M., Elshennawy, N.M.: Classification of brain MRI tumor images based on deep learning PGGAN augmentation. *Diagnostics (Basel, Switzerland)* **11**(12), 2343 (2021). <https://doi.org/10.3390/diagnostics11122343>
3. Husin, N.A., Husam, M., Hussin, M.: Classification of brain tumors: using deep transfer learning. *J. Theor. Appl. Inf. Technol.* **101**(1), 223–235 (2023)
4. Xiaojian, W.: Application of 1.5TMRI Diffusion tensor imaging technology in the diagnosis of brain malignant tumors. *Foreign Lang. Ed. Med. Health* **1**, 194–196 (2021)
5. Ilaria, L., et al.: QOL-35. Analysis on the evolutionary pattern of cognitive functions in children on follow-up for brain tumors. *Neuro Oncol.* **26**(Supplement_4), 0 (2024). <https://doi.org/10.1093/neuonc/noae064.623>
6. Ayadi, W., Elhamzi, W., Atri, M.: Multi-classification of brain tumor based on deep CNN. In: 2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 87–90 (2022). <https://doi.org/10.1109/SETIT54465.2022.9875468>
7. Agalya, V., et al.: CPRO: competitive poor and rich optimizer-enabled deep learning model and holoentropy weighted-power K-means clustering for brain tumor classification using MRI. *Int. J. Pattern Recognit Artif Intell.* **4**, 36 (2022). <https://doi.org/10.1142/S0218001422520085>
8. Brima, Y., et al.: Deep Transfer Learning for Brain Magnetic Resonance Image Multi-class Classification (2021). <https://doi.org/10.48550/arXiv.2106.07333>
9. Alaraimi, S., Okedu, K.E., Tianfield, H., Holden, R., Uthmani, O.: Transfer learning networks with skip connections for classification of brain tumors. *Int. J. Imaging Syst. Technol.* **31**, 1564–1582 (2021)
10. Shunmugavel, G., Suriyan, K., Arumugam, J.: Magnetic resonance imaging images based brain tumor extraction, segmentation and detection using convolutional neural network and VGC 16 model. *Am. J. Clin. Oncol.* **47**(7), 339–349 (2024). <https://doi.org/10.1097/COC.0000000000001097>
11. Babu, K., et al.: An effective brain tumor detection from T1w MR images using active contour segmentation techniques. *J. Phys. Conf. Ser.* **1804**(1), 012174 (2021). <https://doi.org/10.1088/1742-6596/1804/1/012174>



Construction and Experimental Verification of Automatic Classification Process Based on K-Mer Frequency Statistics

Pengwei Zhu^(✉)

University of Texas Health Science Center at Houston, Houston, TX 77054, USA

Pengwei_James_Zhu@163.com

Abstract. In bioinformatics, k-mer frequency statistics are an important tool for analyzing biological sequences, widely used for sequence alignment, duplicate detection, correction, species identification, and motif discovery. However, when dealing with large-scale data, existing k-mer frequency statistics tools have significant bottlenecks in the use of memory and disk space. Therefore, this article proposes a new k-mer frequency statistical method - KCOSS, aimed at optimizing storage and computing efficiency. When processing sequences with a length not exceeding 14, KCOSS uses static hash tables and composite bijective functions, effectively reducing storage requirements. For sequences longer than 14, KCOSS combines Bloom filters and two-level hash tables (including static hash tables and dynamic cuckoo hash tables) to improve processing speed. In addition, to further reduce memory and disk usage, KCOSS optimizes continuous k-mers by only storing newly emerging bases. This article also implements a lockless thread pool, a lockless segmented Bloom filter, and a lockless compact hash table to reduce competition for shared memory and improve parallel processing performance. The experimental results show that when processing human genome data, KCOSS is 22.91% to 169.90% faster than KMC3 under 24 threads, and 527.62% to 806.43% faster than Jellyfish 2; Under 48 threads, the speed improvement ranges from 77.27% to 170.58% and 529.60% to 675.84%, respectively. In terms of memory consumption, KCOSS is comparable to KMC3, only 16.67% of Jellyfish 2; In terms of hard disk storage, KCOSS requires only 12.40% to 17.84% of Jellyfish 2's space and 15.73% to 30.69% of KMC3's.

Keywords: Bioinformatics · K-Mer Frequency Statistics · Parallel Optimization · Genomic Data Analysis

1 Introduction

Since the completion of the Human Genome Project (HGP) in April 2003, the rapid advancement of high-throughput sequencing technology has significantly reduced the cost of genome sequencing, by approximately 50000 times. This change enables researchers to obtain larger scale data in a shorter amount of time. Nowadays, a single laboratory can process TB or even PB level data at a lower cost, and this amount of

data has grown thousands of times between 2000 and 2010. Entering the ‘post genomic era’, life science research is gradually shifting from basic data acquisition to data sharing and in-depth analysis. Various types of bioinformatics data are widely uploaded to databases, allowing researchers to quickly access these data and conduct comprehensive research by combining technologies from different fields. Bioinformatics, as an interdisciplinary field that integrates biology, applied mathematics, statistics, information science, and computer science, is rapidly developing. In bioinformatics, biological sequence data analysis is a crucial field aimed at extracting valuable information from large amounts of data, revealing biological features, decoding genetic information, and ultimately explaining life phenomena. At the same time, multi-core computing technology has rapidly developed due to its advantages in power consumption, heat generation, and scalability compared to single core computing. Multi core computers can perform more tasks simultaneously when processing large-scale biological sequence data, which demonstrates significant advantages in applications such as sequence analysis and comparison. In biological sequence data analysis, k-mer frequency statistics provide key sequence feature information, which helps to reveal the similarities between sequences and genetic differences among different organisms. As a fundamental problem in bioinformatics, k-mer frequency statistics play a central role in tasks such as genome assembly, multiple sequence alignment, duplicate detection, sequence correction, species identification, mutation detection, motif discovery, and sequence correction. Given the importance of k-mer frequency statistics and the advantages of multi-core computing, adopting parallel technology to accelerate this process has significant practical significance.

2 Related Research

2.1 Biological k-mer Sequence

In bioinformatics, k-mer is a fundamental and crucial concept. K-mer refers to a continuous subsequence of length k extracted from a biological sequence. This concept is widely used in various biological data analysis tasks, such as genome assembly, sequence alignment, mutation detection, and species identification. YZ Zhang and colleagues compared the learned k-mer embeddings with commonly used k-mer representations in sequence-based function prediction tasks [1]. They also proposed a novel solution to accelerate the pre-training process. M Ravikumar and colleagues employed machine learning algorithms based on k-mer functions to classify DNA sequences. They effectively retrieved matching sequences using pattern-matching algorithms [2]. The results demonstrated that the SVM linear classifier performed well in this classification task. D Marrama and colleagues developed a new tool called PEPMatch in their study. This tool uses a deterministic k-mer mapping algorithm to preprocess protein datasets, achieving a 50-fold speed increase over conventional local alignment search tools (such as BLAST) without compromising recall [3]. Z Teng and his team introduced a novel method called MFSLNC in their study. This approach measures the functional similarity of lncRNAs using variable k-mer spectra. MFSLNC employs a dictionary tree for storage, allowing for comprehensive representation of lncRNAs with long k-mers [4].

2.2 Overlapping Sequences

C Wei and his team proposed a new method for predicting protein-coding regions in transcriptional sequences, based on a bidirectional recurrent neural network with non-overlapping trimer features [5]. This method has shown significant improvements compared to existing approaches, but there remains considerable potential for further performance enhancements. Petti and SR Eddy described two novel methods for splitting sequence data into distinct training and testing sets. These algorithms take a sequence family as input and produce a split in which the similarity between each test sequence and any single training sequence is less than $p\%$ [6]. These methods have successfully split more families than previous approaches, enabling the creation of more diverse benchmark datasets. MM Breve and his team introduced a novel approach for classifying biological sequences by integrating complex network analysis with entropy maximization. This method applies the principle of maximum entropy to identify the most informative edges related to RNA categories, thereby creating a filtered complex network [7].

2.3 Statistical Model

In their study, M Sedighi and colleagues used linear mixed models (LMM) to analyze the longitudinal changes in clinical indicators of patients [8]. They incorporated gender and age as covariates in the LMM to examine their associations with adverse events (AE) and clinical metrics. A Buratin and colleagues discussed using a generalized linear mixed model (GLMM) in their research to analyze circRNA abundance count data from multiple tools within a single framework [9]. This model accounts for the correlation structure both within and between the quantification tools. In their paper, T Lin and W Wang developed a generalized multivariate linear mixed model that accommodates both censored responses and non-negligible missing outcomes [10]. They introduced a computationally feasible Monte Carlo Expectation Conditional Maximization (MCMC) algorithm for parameter estimation using maximum likelihood (ML) methods. Additionally, they proposed a general information-based approach to assess the variability of ML estimators.

3 Method

3.1 Overall Framework of the Plan

The KCOSS frequency counting method involves several key stages: data input, data preprocessing, frequency counting, data output, and data conversion. Figure 1 illustrates the complete workflow of the KCOSS scheme.

The k-mer frequency statistics scheme of KCOSS adopts different processing methods based on the different values of k . When the value of k does not exceed 14, the system generates a k-mer frequency table containing location information, which can effectively save additional storage space. For cases where the k value exceeds 14, users can choose from three output forms based on their specific analysis needs: if they need to analyze non single occurrence k-mers, they can choose to output these non single instance k-mers, although Bloom filters may introduce a small number of false positives, causing

k-mers with a frequency of 1 to occasionally be mistakenly included. In order to obtain comprehensive k-mer frequency statistics, it is possible to choose to output complete data including overlapping sequence sets and non singleton k-mers. This way, even if some k-mers are misclassified due to the misjudgment of the Bloom filter, the final statistical results will still remain accurate and only unnecessary space waste will occur. If you need to check the existence and frequency of a specific k-mer, you can choose to output in the form of a Bloom filter bit set and a non singleton k-mer. This method preserves the bit set and non singleton k-mer of the Bloom filter, and stores them through hash mapping, thereby achieving efficient space utilization and fast querying. Its space consumption is not affected by the value of k, only by the number of hash functions used, and the false positive rate depends on the setting of the Bloom filter. Dump conversion is similar to the inverse operation of frequency statistics. Its purpose is to derive the key value pair information of the original k-mer sequence (applicable for $k \leq 14$) by parsing binary k-mer frequency data and its position information, or to re parse and merge the overlapping sequence set containing complete sequence information with non singleton k-mer binary files, ultimately generating a text key value pair form containing k-mer sequence information and statistical results.

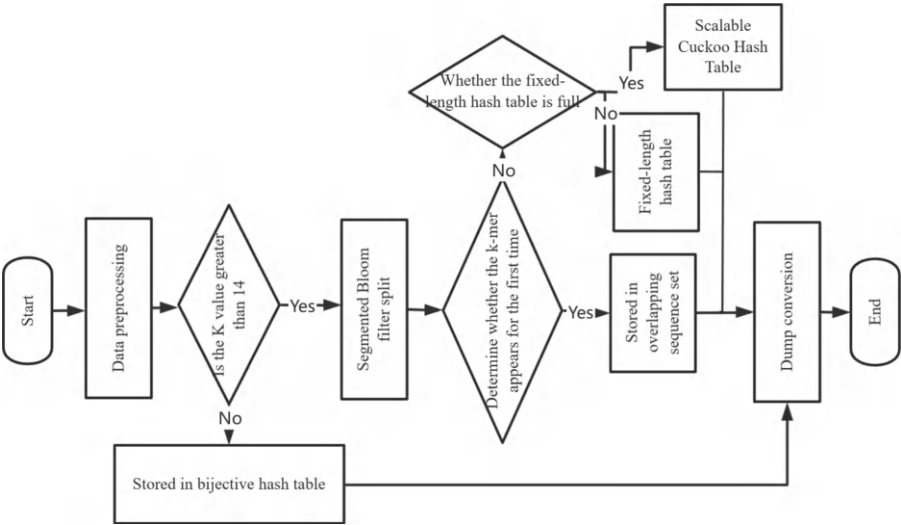


Fig. 1. Overall process of k-mer frequency statistics plan

3.2 Construction of Overlapping Sequence Sets

The overlapping sequence set consists of multiple long segment sequences that overlap and connect with each other through the same segment sequence. These long segments with significant sequence redundancy are called overlapping sequence blocks. In order to efficiently utilize memory and hard disk storage space while improving read and write speeds, KCOSS chose to implement its own 32-bit unsigned integer data storage method,

without using existing containers or serialization frameworks in the STL library. In practical design, the set of overlapping sequences is organized within a continuous space in memory, with overlapping sequence blocks arranged closely in sequence. Figure 2 shows the overall layout of this data structure. Each overlapping sequence block includes a flag block (shown in red) and a binary block for storing the sequence (shown in blue). Both parts use 32-bit unsigned integers for byte alignment. The flag block is further divided into a part (m_1) that records the remaining space state of the overlapping sequence block and a part (m_2) that records the length state of the block.

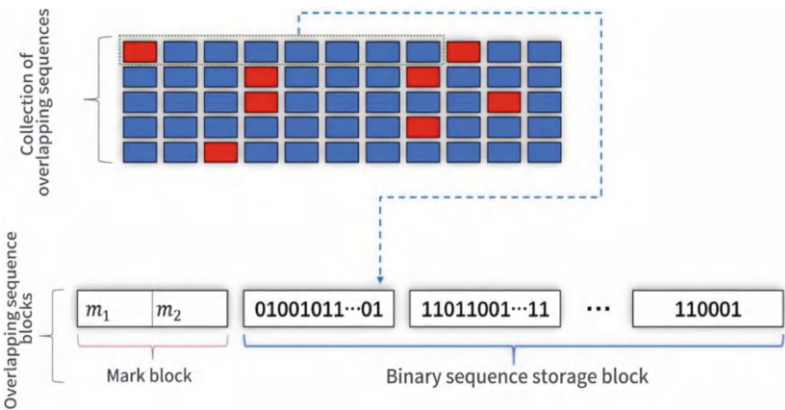


Fig. 2. Schematic diagram of overlapping sequence set data structure

For the case of $k > 14$, this study adopts an innovative processing method: in this scheme, the sequence segments to be processed are called Read, and these sequence segments are segmented by rows. The first occurrence of binary k -mers is called the first k -mer, while subsequent occurrences are non first k -mers. Starting from the first k -mer, long segment sequences formed by overlapping and connecting the same segment sequences are collectively referred to as C-reads. Figure 3 shows the relationship between Reads, C-reads, and k -mers.

Divide it into several k -mer sequences using Read as the unit. For the first appearance of k -mer A, store it in a newly created overlapping sequence block according to the serialization protocol. If the next k -mer of A is also the first k -mer, and because there is only one base difference between A and B, they belong to the same C-reads, then the last base of B can be merged into an existing overlapping sequence block with only 2 storage spaces. If the next k -mer is not the first k -mer, the concatenation of the current overlapping sequence block is ended and stored in a fixed size static hash table. If the hash table is full or reaches the maximum number of probes, store the k -mer in a variable length cuckoo hash table.



Fig. 3. C-reads schematic diagram

3.3 Optimization of k-mer Frequency Statistics Performance Based on Overlapping Sets

The analysis in the frequency statistics stage shows that the processing flow is consistent for each input Read. By using a divide and conquer strategy, the dataset can be split and allocated to multiple threads for parallel processing, thereby improving the throughput of KCOSS. As shown in Fig. 4, KCOSS adopts a thread pool based on the producer consumer model to process sequence data, and evenly distributes the split Read sequence to different threads. To ensure that multithreading can accurately obtain the diversion status from the Bloom filter when $k > 14$, KCOSS uses thread safe Bloom filters. In addition, non singleton k-mer storage also uses concurrent hash tables to ensure the accuracy of statistical results.

When processing k-mer frequency statistics in a multi-threaded environment, multiple threads share the same hash table, and update operations require special attention. Specifically, the thread first uses open addressing to locate the key value pairs of k-mers. If the search exceeds the preset number of times, the thread will turn to a small hash table that supports dynamic expansion for updates. For non-existent key value pairs, the thread will attempt to insert new entries; For existing key value pairs, perform a self-increment operation. In computer systems, threads cache accessed variables to local caches, and these changes are usually not immediately reflected in main memory. This will lead to two main issues:

1. The modification of a hash table by a thread may not be immediately detected by other threads, resulting in the possibility of overwriting existing key value pairs and generating incorrect data.

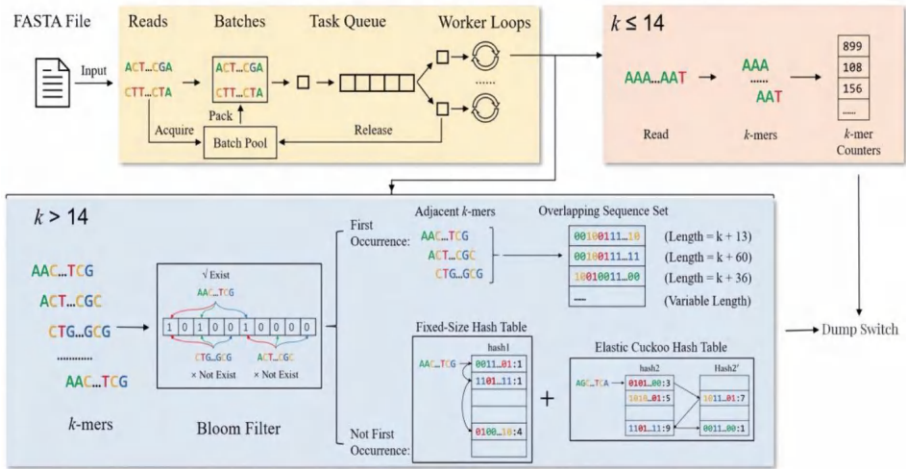


Fig. 4. Frequency statistics process under multithreading

2. Self increment operation is not an atomic operation, it includes reading values from memory, updating values from cache, and then writing back to memory. When multiple threads perform these operations simultaneously, it may result in reading outdated values and writing incorrect data to memory.

To ensure data consistency, mutex locks are generally introduced to ensure that only one thread can operate shared resources at a time. However, this method may incur performance overhead. To improve efficiency, atomic operations can be used for lock free updates, ensuring data correctness even if multiple threads operate in parallel. Specifically, threads first use open addressing to find key value pairs, and if the search fails, update them using the cuckoo hash table; If the key value pair does not exist, use atomic CAS operation to attempt insertion. If it fails, retry; If a key value pair exists, use the atomic fetch add operation for self increment.

4 Results and Discussion

In the experiments, this study employed reference genome sequence data and raw sequencing data from the National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov/>) to evaluate the k-mer frequency counting performance for both assembled datasets and standard sequencing data.

From Fig. 5, it can be seen that as the number of threads increases, the acceleration effect of the KCOSS algorithm gradually strengthens. However, when the number of threads exceeds 40, regardless of whether the k value is 32 or 64, there is a marginal decrease in the acceleration effect. This is mainly because KCOSS uses shared Bloom filters for k-mer splitting. As the number of threads increases, there will be resource competition between threads. Although Bloom filters have achieved lock free optimization through CAS operations, too many threads increase the probability of write failures, which in turn affects the effectiveness of CAS locks. The result is that the thread falls

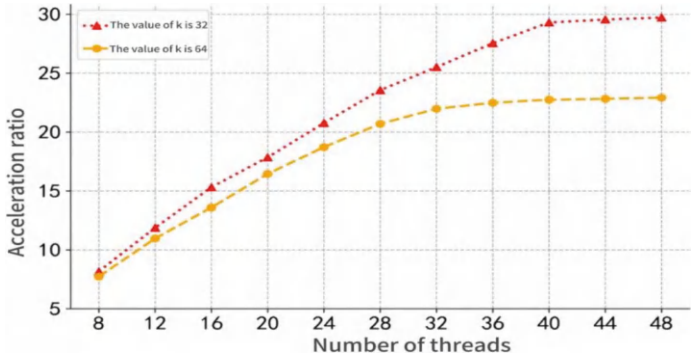


Fig. 5. Experimental study on parallel acceleration ratio of KCOSS algorithm under different k values

into a state of continuous retry during the diversion phase, unable to effectively complete subsequent tasks, and the running time no longer significantly decreases with the increase of the number of threads.

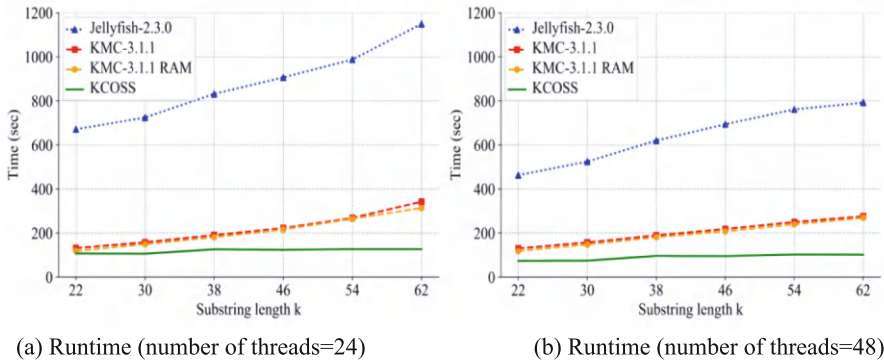
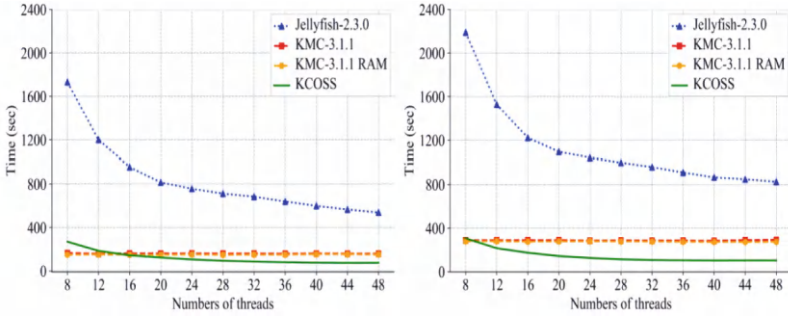


Fig. 6. Comparison of Jellyfish 2, KMC3, KMC3 memory version, and KCOSS runtime under different k values

The experimental results show that KCOSS exhibits excellent performance in processing assembled human genome data. In tests using the GRCh38.p12 dataset, when 24 threads were used, KCOSS was 22.91% to 169.90% faster than KMC3 and 527.62% to 806.43% faster than Jellyfish2. Under the configuration of 48 threads, KCOSS's speed has increased by 77.27% to 170.58% compared to KMC3, while it is 529.60% to 675.84% faster than Jellyfish 2. For the case of a k value of 62, the memory consumption of KCOSS is comparable to KMC3, but only 16.67% of Jellyfish 2's. In terms of hard disk storage, when the k value exceeds 14, KCOSS only requires 12.40% to 17.84% of Jellyfish 2's storage space and 15.73% to 30.69% of KMC3's storage space.

From Fig. 6, it can be seen that when using the human genome GRCh38.p13 as the test dataset and the k value increases from 22 to 62 (with an interval of 8), KCOSS is significantly better than Jellyfish 2, KMC3, and KMC Lam. In the case of multi-threaded



(a) Run time (k value=32)

(b) Run time (k value=64)

Fig. 7. Comparison of Jellyfish 2, KMC3, KMC3 in memory version, and KCOSS runtime under different threads

operation, KCOSS is 22.91% to 170.58% faster than KMC3 and 527.62% to 806.43% faster than Jellyfish 2. It is worth noting that the running time of KCOSS is almost not significantly affected by changes in k value, while the running time of Jellyfish2 and KMC3 increases proportionally with k value. KCOSS utilizes advanced data structures and multi-threaded algorithms, including lock free thread pools, shared hash tables, lock free queues, and cuckoo hash tables. When the k value does not exceed 14, using a shared hash table effectively reduces the cost of contention access and memory consumption. However, as the value of k increases, although independent hash tables improve parallelism, they also lead to significant memory consumption and additional statistical information merging overhead. The lock free queue is used for building lock free thread pools and memory block recycling, avoiding frequent system memory operations.

Figure 7 shows that Jellyfish2 exhibits better acceleration performance when increasing the number of threads, while KMC3 has a smaller response to the increase in thread count. After exceeding 32 threads, the running time of KCOSS tends to stabilize. Despite 48 threads, Jellyfish 2 still requires 544.42% to 789.93% of KCOSS time. The performance improvement of Jellyfish 2 mainly comes from unlocked hash tables and prefix arrays, but the limitation of shared data structures gradually weakens the acceleration effect. Although KCOSS is also subject to similar limitations, due to limited room for improvement, its acceleration ratio remains stable at 32 threads, while Jellyfish 2 still shows a slight improvement at 48 threads.

5 Conclusion

This article proposes and implements a new k -mer frequency statistics scheme called KCOSS, which improves upon the shortcomings of existing tools in terms of memory and disk space consumption. By analyzing the assembled sequence data, we found that when the k value is large, the proportion of single occurrence k -mers usually exceeds 90% of the total k -mer types, revealing a large amount of redundant information. KCOSS adopts two statistical strategies based on sequence length: for short sequences (length ≤ 14), static hash tables and composite bijective functions are used to reduce storage requirements; For long sequences (length > 14), combining Bloom filters and two-level

hash tables significantly reduces storage space requirements. KCOSS has also introduced lock free thread pools and lock free data structures to reduce computational overhead and improve concurrency performance. The experimental results show that compared with Jellyfish2 and KMC3, KCOSS performs well in both speed and storage efficiency. In tests using human genome data, KCOSS is significantly faster than Jellyfish2 and KMC3, and also has more advantages in storage. However, KCOSS exhibits certain limitations in handling shared Bloom filters, particularly in terms of performance at high line counts. In addition, for sequencing sequences, the time efficiency of KCOSS still needs to be improved, and the memory consumption is relatively high. Future work will focus on optimizing the performance of Bloom filters and exploring disk space based statistical algorithms to further improve the efficiency of k-mer frequency statistics under memory limitations.

References

1. Zhang, Y.Z., Bai, Z., Imoto, S.: Dysfunctional analysis of the pre-training model on nucleotide sequences and the evaluation of different k-mer embeddings. *bioRxiv* **37**, 2112 (2022). <https://doi.org/10.1101/2022.12.05.518770>
2. Ravikumar, M., Prashanth, M.C., Guru, D.S.: Matching Pattern in DNA Sequences Using Machine Learning Approach Based on K-Mer Function. (2022). https://doi.org/10.1007/978-3-030-96634-8_14
3. Marrama, D., et al.: PEPMatch: a tool to identify short peptide sequence matches in large sets of proteins. *BMC Bioinform.* (2023). <https://doi.org/10.1186/s12859-023-05606-4>
4. Teng, Z., et al.: Measuring functional similarity of lncRNAs based on variable K-mer profiles of nucleotide sequences. *Methods* (2023). <https://doi.org/10.1016/j.ymeth.2023.02.009>
5. Wei, C., Zhang, J., Yuan, X.: Enhancing the prediction of protein coding regions in biological sequence via a deep learning framework with hybrid encoding. *Digit. Signal Process.* **123**, 103430 (2022). <https://doi.org/10.1016/j.dsp.2022.103430>
6. Petti, S., Eddy, S.R.: Constructing benchmark test sets for biological sequence analysis using independent set algorithms. *PLoS Comput. Biol.* **18**(3), e1009492 (2022). <https://doi.org/10.1371/journal.pcbi.1009492>
7. Breve, M.M., Pimenta-Zanon, M.H., Lopes, Fabrício Martins: BASiNETEntropy: An Alignment-Free Method for Classification of Biological Sequences Through Complex Networks and Entropy Maximization (2022). <https://doi.org/10.48550/arXiv.2203.15635>
8. Sedighi, M., et al.: Linear mixed model analysis to evaluate correlations between remdesivir adverse effects with age and gender of patients with mild Covid-19 pneumonia. *J. Med. Virol.* **94**(8), 3783–3790 (2022). <https://doi.org/10.1002/jmv.27800>
9. Buratin, A., et al.: Detecting differentially expressed circular RNAs from multiple quantification methods using a generalized linear mixed model. *Comput. Struct. Biotechnol. J.* (2022). <https://doi.org/10.1016/j.csbj.2022.05.026>
10. Lin, T.-I., Wang, W.-L.: Multivariate linear mixed models with censored and nonignorable missing outcomes, with application to AIDS studies. *Biom. J.* **64**(7), 1325–1339 (2022)



A Strategy to Determine Priorities Among Multiple Goals: Approaches from Network Models

Mucun Xie¹(✉), Dachao Shang¹, and Shuyan Zeng²

¹ School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
15565990229@163.com

² Department of Geography, Beijing Normal University, Beijing 100875, China

Abstract. UN staff face a difficult decision when they need to prioritize the 17 Sustainable Development Goals (SDGs). Faced with multiple constraints such as financial and human resources, the UN needs to make decisions by building an effective model. Our team developed the W17 relationship network model, the priority assessment model, and other models based on the W17 relationship network, and applied them to prioritize the goals of other companies and organizations. We discuss how the network of relationships changes under the assumption that no poverty goal is achieved, and obtain new priorities for the SDGs by reassigning values to indicators. We considered the impact of international crises on the network structure, discussing the impact of global pandemics, refugee movements, and climate change on the choice of priorities. The impact of these international crises on the progress of the UN's work was analyzed from a network perspective. Finally, we applied the relationship network construction model and the priority assessment model to prioritize the goals of other companies and organizations. We also performed a sensitivity analysis of the model. The percentage of synergy still accounts for more than 80% of our artificially determined indicator increments when they fluctuate in the 2% interval. This indicates that our model is stable within a reasonable range of variation.

Keywords: Network Security · Sustainable Development Goals · Relationship Networks · Priorities · Correlation Factors · Synergies

1 Introduction

When there are multiple goals under the overall goal, the question becomes how to prioritize among them, especially when the goals are somehow linked and not independent of each other. The 17 Sustainable Development Goals (SDGs) set by the United Nations are a good example [1]. These goals are not relatively independent of each other, for example, SDG13 (climate action) could potentially have a negative impact on SDG8, while SDG4 (quality education) could help achieve SDG1 (no poverty) [2]. This poses a challenge for the UN to complete action on the SDGs. It has been more than five years since the UN released the SDGs, and the world has changed a lot [3]. Many old

issues have not been well addressed, such as the fact that UN funding remains somewhat limited, the international situation is still volatile, and poverty is still widespread. New events happened [4], including the improvement of the EU Emission Trading System (EU ETS) and the global pandemic [5], all of which have an impact on the achievement of the Sustainable Development Goals. At this important point, it is critical for the UN to take effective action to find a comprehensive solution that will help prioritize these 17 goals as soon as possible. In order to study and prioritize the interrelationships among the 17 Sustainable Development Goals (SDGs) set by the United Nations and contribute to the achievement of the SDGs [6], our team developed the W17 relationship network construction model, the priority assessment model, and other models based on the W17 relationship network, and applied them to other companies' and organizations' goal prioritization. The specific tasks we accomplished are as follows:

- Developed a model that articulates the network relationships among the 17 Sustainable Development Goals (SDGs).
- A priority assessment model was developed to obtain the priorities of the 17 SDGs.
- The changes in the network of relationships under the assumption that no poverty goal is achieved are discussed, and it is suggested that the United Nations include "low-cost pollution control technologies" in the SDGs.
- Reflects on the impact of global pandemics, refugee movements and climate change on the structure of the network and the work of the UN.
- Applied the relationship network building model and priority assessment model to prioritize goals for other companies and organizations.

2 W17 Relationship Network Model

2.1 Model Description

In 2015, a plan for achieving sustainable development was jointly developed, consisting of 17 sustainable development goals and 169 targets (hereafter referred to as goals). These Goals relate to different aspects of human life and do not exist in isolation; their relationships are complex and interactive, and the achievement of one Goal may have varying degrees of impact on the ability to achieve others. Therefore, in order to estimate the interrelated system of SDGs, we have applied a network approach to the SDGs, using network theory to determine how the system evolves over time and what its likely outcomes are in the face of internal and external change. We developed a three-layer network structure, called the W17 model. The first layer of the network structure consists of a composite of 17 SDGs, or SDGs as proposed by the UN. These 17 SDGs are composed of a second layer of the network structure consisting of 169 targets, each of which is subdivided into indicators, which form the third layer of the network. These 169 targets are subdivided into indicators, which form the third layer of the network. As shown in Fig. 1.

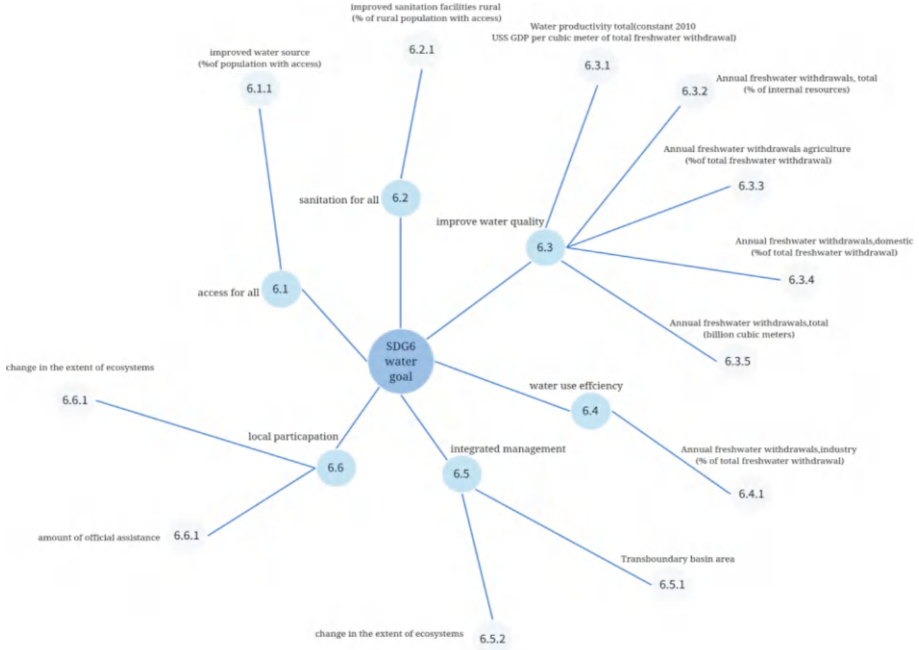


Fig. 1. SDG6 network structure diagram

2.2 Construction Method (Spielman Correlation Coefficient and Adjacency Matrix)

We used 331 indicators from data published by the World Bank to identify the interactions between the 71 targets and the 17 SDGs. When we establish relationships between World Bank indicators, we are indirectly linking the indicators, targets and SDGs [7].

We use Spearman's correlation coefficient (ρ) to estimate the strength of the correlation between two indicators. Unlike Pearson correlation analysis, Spearman correlation analysis captures non-linear correlations between variables, is less sensitive to outliers and is widely used for general relationships other than linear correlations between two indicators. P-values were then calculated and correlations with p-values less than 0.05 were considered statistically significant. The indicators were therefore linked at a significance level of $p < 0.05$.

The network connections between the specific target layer and the SDG layer are represented by their adjacency matrices. The correlations between indicators are stored in the matrix AW (Indicator - Target; 331×331) and targets and indicators can be represented by multiplying the adjacency matrix, the following matrix equation shows the representation of the target layer connection matrix Att (Target - Target; 71×71):

$$A_{tt} = A_{ti}^T \times A_{ij} \times A_{ti} \quad (1)$$

The linkages between the SDGs Agg (SDG - SDG; 17×17) can be represented by the projection of Agt (SDG - Target; 17×71) and Att (Target - Target; 71×71) as:

$$A_{agg} = A_{gt}^T \times A_{tt} \times A_{gt} \quad (2)$$

2.3 Model Result: A Network Diagram

By calculating the matrix Agg, we obtained the magnitude of the interactions between the SDGs. Based on this data, we plotted the network of relationships between these 17 SDGs as follows:



Fig. 2. 17SDGs Network Structure

As shown in Fig. 2, each node corresponds to an SDG, and the size of the node corresponds to the feature vector centrality of that SDG. The blue line indicates positive correlation, the orange line indicates negative correlation, and the thickness of the line indicates the magnitude of the correlation, ranging from -1 to 1 .

3 Priority Assessment Model

Through the construction of the relational network, we derived complex and mutually influential relationships among the 17 SDGs, 169 targets, and 331 indicators. Their implicit interdependencies may lead to conflicting situations among the SDGs, resulting

in different outcomes. By calculating the matrix A_{ii} , we can obtain the interactions among the global SDG indicator pairs. The synergistic, hindering, and irrelevant interactions between the SDG indicator pairs construct the interrelationships among the 17 SDGs. In order to prioritize the SDGs, it is necessary to assess the synergistic effects of each SDG, that is, the ability of each SDG to drive the achievement of other goals. After calculating, the results are shown in the Fig. 3:

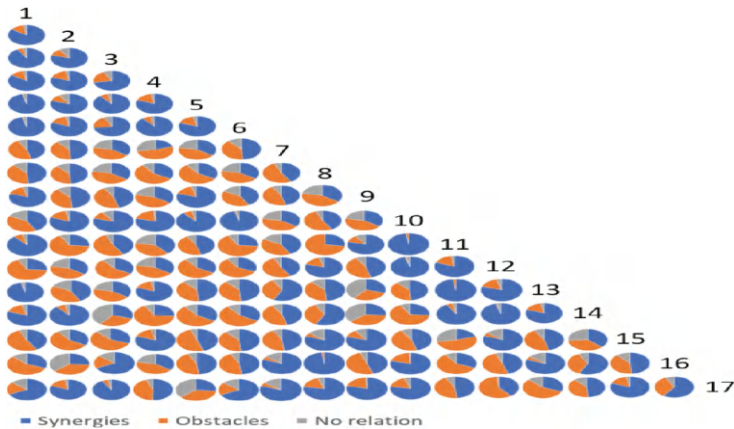


Fig. 3. Percentage graph of three interactions

Analyzing the statistical results, it can be seen that SDG1 (no poverty) has a synergistic relationship with 75% of the SDGs; secondly, SDG3 (good health) also has a large synergistic effect and will also play a key role in contributing to the achievement of the UN goals.

In order to further filter out the higher priorities and make the estimation results more accurate, we turned our attention to the matrix A_{tt} (which must be symmetric according to the actual meaning of the matrix), where the number a_{ij} in the matrix indicates the magnitude of the correlation between SDG $_i$ and SDG $_j$. We ranked and took out the top ten to become the top ten global synergistic relationships, and the results are shown in the following Fig. 4.

The results are analyzed as follows: among the top ten global synergistic relationships, there are five synergistic relationships involved in the construction of SDG1 and three synergistic relationships involved in the construction of SDG3, from the perspective of matrix method, the priority of SDG1 should be higher than that of SDG3. Combining the above two analysis results, we can conclude that the priority is SDG1.

We consider a forecasting model with time series analysis to calculate estimates of these indicators for the next ten years. We selected data for the two indicators for the decade 2010–2019 because in 2020–2022 there was a sudden global pandemic that had a significant impact on economic development, so the data for these two years had a large change from the previous trends, which we can consider as statistical outliers and are not counted in our analysis. Applying a time series (ARMA model) to the analysis,



Fig. 4. Top 10 synergy effect pairs

the

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right)(1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t \tag{3}$$

We obtain the following forecast curves and data.

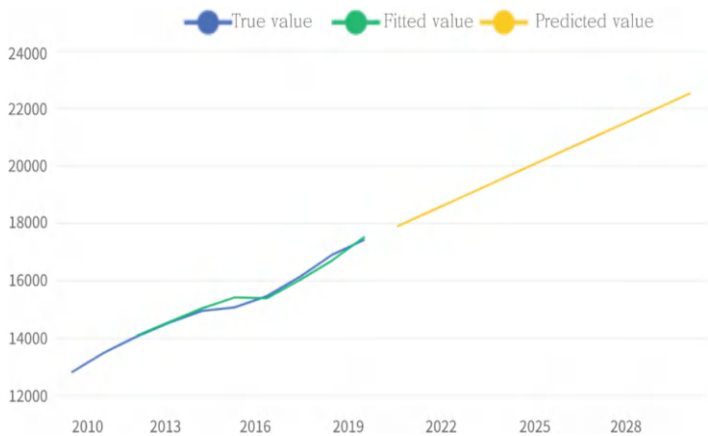


Fig. 5. GNIPC Prediction figure

It can be seen from the analysis of Figs. 5, 6, and 7, it is clear that if priority SDG1 is activated, the total number of unemployed people in the world will decrease by 10.2454% of the total labor force in the next ten years, while the per capita national income will increase by 25.9267%. Then we can reasonably assume that other indicators that are more correlated with these two indicators, such as “higher education school enrollment rate” and “service employment rate”, will also be improved to some extent, and then

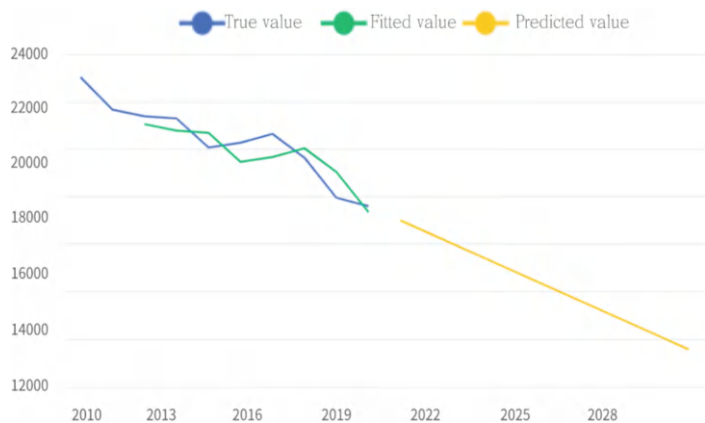


Fig. 6. Prediction figure of the number of unemployed people as a percentage

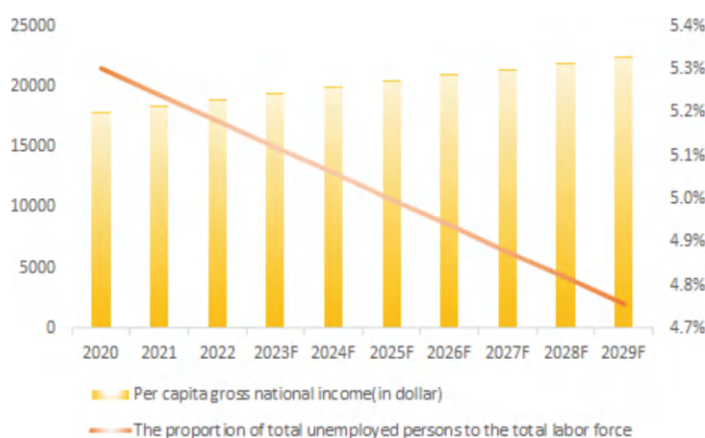


Fig. 7. Unemployment Share and Income Prediction Table

they will be positively correlated with SDG will be improved to different degrees, i.e., SDG4 and SDG8 will be reasonably achieved.

4 Network Data Collection and Big Data Analysis

The network model we constructed represents the state of interactions and interactions between indicators and between specific targets, but we prefer to understand how these interactions affect the evolution of this network toward the ultimate goal of overall sustainability. To estimate this, we used the graph Laplacian matrix algorithm. To estimate this, we used the graph Laplacian matrix algorithm. The algorithm is used to determine the stability of the network using the relational network matrix Agg (we use AG below for the convenience of writing later) the graph Laplacian matrix LP (the network can be

considered stable when all eigenvalues of the graph Laplacian matrix LP are ≤ 0) and the algorithm for the matrix is as follows:

$$L_{ij} = \begin{cases} A_{ij}, & i \neq j \\ -\sum_k A_{ij}, & i = j \end{cases} \quad (4)$$

The positive eigenvalues obtained from the graph Laplace decomposition results indicate that the network is unstable in achieving SDGs, and will change with the achievement of one SDG.

We construct a “proportional indicator growth function” to calculate the magnitude of indicator changes in order to predict changes in network relationships. The independent variable of this function is the correlation coefficient between indicator A and indicator B in the SDG1 network branch for the 16 SDGs except SDG1, and the dependent variable is the change in indicator A. The relationship between the independent variable and the dependent variable is a positive proportional function with a coefficient of 5%. Since the indicators are not simply connected to each other, there may be synergistic or obstructive relationships between indicator A and multiple indicators in the SDG1 network branch, therefore, after using the “indicator data growth proportional function” to calculate the change of indicator A affected by multiple indicators in the SDG1 network branch, the change should be The final result is the magnitude of change in indicator A when the SDG1 goal of no poverty is achieved. The calculation formula is as follows.

$$y = \frac{1}{n} \sum_{i=0}^n 5\% \times \rho \quad (5)$$

where n is the number of indicators related to the indicator.

Based on the new network relationships, we again use the Optimization Matters Assessment Model to explore the impact of achieving the SDG without poverty on the priorities. In the end, we arrive at the new priority ranking as follows.



Fig. 8. 17SDGs Priority Pyramid chart

As shown in the Figs. 8 and 9, the priorities of the remaining 16 SDGs change from before the implementation of SDG1: SDG3 becomes a priority; SDG2, SDG3, SDG6, SDG11, SDG14, SDG8, SDG9, SDG15, and SDG17 decrease in priority; SDG5, SDG8, SDG10, SDG12, SDG13, and SDG14 increase in priority.

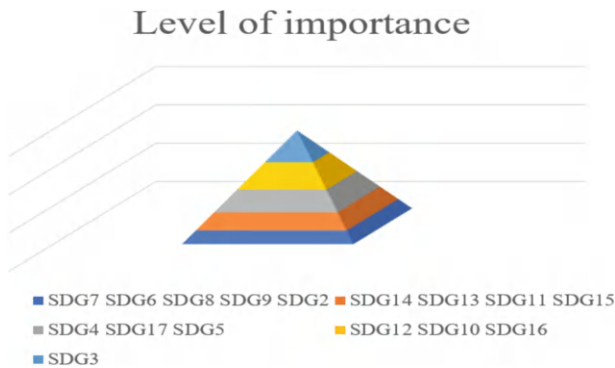


Fig. 9. 16SDGs Priority Pyramid chart

5 Model Analysis

In the third question, we constructed a function to calculate the increment of the indicator, where we defined this rate of change as a fixed value of 5%. We want to understand whether the magnitude of this rate of change would have a significant impact on the results, within a reasonable range. The results were expressed in terms of the percentage of synergy observed between SDG3 and the other SDGs (a larger percentage represents a stronger priority). Holding all other data constant, we varied the magnitude of the rate of change and observed the following results:



Fig. 10. Sensitivity analysis

From this Fig. 10, we can see that in the range of 3% to 8%, the percentage of our synergy is about 80%, and the results in this interval do not change significantly,

the sensitivity is low, and the model is more stable, indicating that our rate of change between 3% and 8% is reasonable, while the rate of change between 1% and 2%, and the deviation of the results at 9% to 10% are relatively large, and we can also easily analyze the reasons. If the data change is not large, the resulting new network model may be similar to the previous network model, and no reasonable conclusion can be drawn. And if the data change is too large, the impact of SDG1 implementation may be too much amplified, resulting in too high a priority for SDG3, which in turn will affect our network model as a whole and lead to wrong judgments when estimating other SDGs. Therefore, we conclude that when the variation rate is between 3% and 8%, the model is more stable and the results obtained have high confidence, and too high or too low variation rate will lead to failure or large deviation in the evaluation of the model.

6 Conclusion

To explore the linkages among the 17 SDGs set by the United Nations [8], we established the W17 relational network model to construct their relational networks. We linked the 17 SDGs, 71 targets, and 331 indicators by calculating the Spearman coefficient and the adjacency matrix to construct a three-level network structure [9, 10]. We propose a priority assessment model, which can assess the dependency of potential interactions among targets and identify the priorities among the 17 SDGs. We then analyze, from a network perspective, what could be achieved in the UN's work over the next decade if the priorities were activated. We then analyze the changes in the network assuming that the goal of no poverty is achieved, and propose a new goal for the UN to include: low-cost pollution control technologies. In addition, we apply the network construction model and the priority assessment model to the goal prioritization of other companies and organizations, which will help companies and organizations to advance their goal achievement process. Finally, we conduct a sensitivity analysis of the model and discuss the strengths and weaknesses of the model, again demonstrating the credibility of our model.

References

1. Gyimah, P., Appiah, K.O., Appiagyeyi, K.: Seven years of United Nations' sustainable development goals in Africa: a bibliometric and systematic methodological review. *J. Clean. Prod.* **395**, 136422.1-136422.11 (2023)
2. Philip, D.C.: The role of exercise physiology in the United Nations' sustainable development goals. *Eur. J. Appl. Physiol.* (2023). <https://doi.org/10.1007/s00421-023-05180-w>
3. Goubran, S., et al.: Green building standards and the United Nations' sustainable development goals. *J. Environ. Manag.* **326**(Pt A), 116552 (2023). <https://doi.org/10.1016/j.jenvman.2022.116552>
4. Scanlon, A., et al.: United Nations' sustainable development goal 3 target indicators: examples of advanced practice nurses' actions. *J. Nurse Pract.* (2022). <https://doi.org/10.1016/j.nurpra.2022.03.005>
5. Shabbir, M.: Exploring the relationship between sustainable entrepreneurship and the United Nations sustainable development goals: a comprehensive literature review. *Sustain. Dev.* (2023). <https://doi.org/10.1002/sd.2570>

6. Anouti, A., Chaperon, S., Kennell, J.: Tourism policy and United Nations sustainable development goal 16: peace and stability in the Middle East and North Africa. *Worldw. Hosp. Tour. Themes* **15**(2), 108–116 (2023). <https://doi.org/10.1108/WHATT-10-2022-0115>
7. The World Bank, Dataset of All Indicators (2018). <https://data.worldbank.org/indicator?tab=all>. Accessed 6 Jan 2019
8. Mukonza, S.S., Chiang, J.L.: Meta-Analysis of Satellite Observations for United Nations Sustainable Development Goals: Exploring the Potential of Machine Learning for Water Quality Monitoring (2023)
9. Umar, T., Umeokafor, N.: Exploring the GCC progress towards united nations sustainable development goals. *Int. J. Soc. Ecol. Sustain. Dev. (IJSESD)* **13**, 1 (2022)
10. Kashnitsky, Y., et al.: Identifying Research Supporting the United Nations Sustainable Development Goals (2022). arXiv e-prints



Innovative Application of Bayesian Algorithm in Network Security Risk Assessment Model

Haosheng Li^(✉), Qingqing Ren, Wei Chen, Yixuan Ma, Qingwang Zhang,
and Wanting Lv

State Grid XinJiang Information and Telecommunication Company, Urumqi 830063, Xinjiang,
China

acain5810@gmail.com

Abstract. With the rapid development of the Internet and digital technology, network security issues are increasingly being taken seriously, and the frequency and complexity of attacks are also increasing. This article focuses on the innovative application of Bayesian algorithm in the field of network security risk assessment, aiming to construct a highly adaptable dynamic risk assessment framework to cope with the complex and uncertain network environment. This article delves into the unique advantages of Bayesian networks in integrating diverse heterogeneous data, including network traffic, system logs, vulnerability intelligence, etc., with the aim of achieving comprehensive and in-depth risk assessment. This study achieved early warning and precise analysis of potential network threats by finely constructing a Bayesian network model and applying its powerful reasoning ability. During the experimental phase, we carefully planned a series of realistic network environment simulation scenarios to fully validate the effectiveness, robustness, and generalization ability of the proposed model. The model scores for data points 1 to 3 are all between 0.55 and 0.75, and the adaptability scores are also relatively high (0.60 to 0.80), indicating that on platform A, the model's judgment of normal and abnormal data points is relatively accurate. In summary, the risk assessment model based on Bayesian algorithm proposed in this study has laid a solid technical foundation for building a stable and reliable network security ecosystem.

Keywords: Bayesian Algorithm · Network Security Risk Assessment · Network Threats · Environmental Adaptability

1 Introduction

With the rapid development of information technology, computer networks have become the core network for information flow. However, this has been accompanied by a sharp increase in network security threats, which are unprecedented in complexity and scale, posing severe challenges to the economy and society. Traditional risk assessment methods are limited by static models and find it difficult to capture the rapidly changing network environment, especially rule-based intrusion detection systems, which face high false positives and false negatives due to their lack of flexibility. Currently, although

feature-based machine learning has shown effectiveness in identifying known attacks, it falls short in dealing with unknown threats. In view of this, this study innovatively introduces Bayesian algorithm and utilizes its powerful probabilistic inference mechanism to construct a dynamic network security risk assessment model. This model aims to improve the accuracy and foresight of risk assessment, by dynamically adapting to changes in the network environment, effectively identifying and predicting potential threats, and providing scientific basis for building a more robust network security defense line.

This article focuses on the innovative application of Bayesian algorithm in the field of network security risk assessment, with the core goal of building a risk assessment framework that can flexibly respond to dynamic changes and uncertainties. Deeply explored the potential of Bayesian networks in integrating multi-source data (including network traffic, system logs, and vulnerability information) for comprehensive risk assessment.

This article focuses on the innovative application of Bayesian algorithm in the field of network security risk assessment, with the core goal of building a risk assessment framework that can flexibly respond to dynamic changes and uncertainties. It deeply explores the potential of Bayesian networks in integrating multi-source data (including network traffic, system logs, and vulnerability information) for comprehensive risk assessment. Through the in-depth application of Bayesian inference, this study aims to achieve early insight and accurate analysis of potential network threats. In the experimental stage, we carefully designed various network environment scenarios to comprehensively verify the effectiveness and stability of the proposed model. The research not only expands the theoretical boundaries of network security risk assessment, but also provides novel and practical security analysis tools and strategic perspectives for practical fields. This study is based on a Bayesian algorithm risk assessment model, which will significantly enhance the intelligence and foresight of network security defense, laying a solid foundation for building a more unbreakable network security ecosystem.

In the research framework of this article, firstly, we conduct an in-depth analysis of existing network security risk assessment methods and identified their limitations in handling dynamics and uncertainties. Then, an innovative risk assessment model based on Bayesian networks is proposed, which achieves the fusion and analysis of multi-source data by constructing a probability graph model. Next, we elaborate on the process of building the model, including data collection, model initialization, parameter learning, and the design of inference mechanisms. In the experimental section, the performance of the model is tested in a simulated network environment to analyze its performance in terms of risk identification accuracy and response speed. Finally, we conduct a comprehensive discussion on the experimental results, pointing out the advantages and possible improvement directions of the model, and summarizing the contributions of this paper and the potential for future research. This article aims to explore a new paradigm for network security risk assessment by introducing Bayesian algorithm, laying the foundation for achieving more accurate and real-time network threat detection.

2 Related Work

In the past decade, network security risk assessment has gradually become one of the focus areas of academic and industrial attention. With the diversification and complexity of network attack methods, researchers have developed many risk assessment models

and methods to enhance network protection capabilities. Rule based intrusion detection systems (IDS) are one of the earliest widely used methods, which identify known threats through pre-defined rule sets. Wang Saie proposed a 5G network security risk assessment method based on attack graphs [1]. Zhang Yan proposed a key technology for network security risk assessment based on the concept of hierarchical protection [2]. He Zhengzhang proposed a multimodal transportation network security risk assessment algorithm [3]. Liu Zhen provided research on network security risks and protection measures for rail transit train control systems, and proposed corresponding solutions [4]. Sengupta S believes that due to the static nature of network services and configurations, network defense based on traditional tools, techniques, and programs cannot consider the inherent advantages of attackers [5]. Jain AK believes that with the rapid development of technology, online social networks have become rapidly popular in the past few years. The information shared on social networks and media spreads very quickly, almost instantly, making it very attractive for attackers to obtain information. There are many security and privacy issues related to sharing information with users, especially when users upload personal content such as photos, videos, and audio, attackers can maliciously use the shared information for illegal purposes [6]. However, with the rapid evolution of attack techniques, these systems have shown significant shortcomings in dealing with unknown threats, unable to update rule sets in a timely manner and adapt to new types of attacks. The study also shows that traditional methods have certain limitations in dealing with real-time and dynamic network environments, often resulting in high false alarm and missed alarm rates. These limitations have prompted researchers to explore new methods to improve the accuracy and timeliness of assessments.

In recent years, machine learning methods have gradually been applied in the field of network security, especially in risk assessment and threat detection, and have made certain progress. These methods automatically identify abnormal patterns in network traffic and logs through data-driven approaches, demonstrating strong adaptability. However, existing machine learning methods often rely on large labeled datasets for training and still face challenges in dealing with dynamic changes and uncertain factors. Specifically, machine learning models are susceptible to changes in data distribution and struggle to maintain high accuracy in the absence of sufficient training data. Khan M believes that as network threats become increasingly complex, integrating machine learning techniques into network security has become a top priority for detecting and mitigating the constantly evolving attacks. However, deploying machine learning models in secure applications brings new challenges in the form of adversarial machine learning. Adversarial attacks exploit vulnerabilities in machine learning models, impair their effectiveness, and may lead to security vulnerabilities [7]. Zamani E explored standards and regulations related to blockchain, investigated and analyzed 38 blockchain incidents to determine root causes and provide a view of the most common vulnerabilities [8]. Mughal A A designed a cloud network security architecture for protecting networks in virtual environments [9]. Strida D N proposed a wireless network security analysis and evaluation scheme based on penetration testing execution standards [10]. Sun X categorized various network security risks and vulnerabilities in the automotive industry environment into in vehicle network attacks, vehicle to everything network attacks, and other attacks based on the types of communication networks and attack targets [11]. Somasundaram

R conducted a comprehensive review on the security challenges of medical IoT [12]. In addition, many methods lack transparency and interpretability in the decision-making process, which affects the trustworthiness and acceptance of the model results. Therefore, how to design a risk assessment model that can adapt to dynamic environments and provide reliable explanations remains an important direction and challenge in current research.

3 Method

3.1 Data Collection and Preprocessing

In network security risk assessment, the quality and accuracy of data directly affect the reliability of the assessment results. To achieve this, it is necessary to first collect heterogeneous data from multiple sources, including network traffic data, intrusion detection system logs, system vulnerability information, user behavior records, etc. These data sources are diverse, with different data formats and collection methods for different types of data. Network traffic data can be captured through network monitoring tools such as Wireshark, intrusion detection logs are typically generated by existing security devices, and vulnerability information can be obtained through scanning tools such as Network Mapper.

After the data collection is completed, it is necessary to preprocess these data in order to unify the format and remove noise. The preprocessing process includes data cleaning, data transformation, and data normalization. Data cleaning involves removing duplicate records and erroneous data, data transformation involves format conversion of raw data (such as converting IP addresses to numerical form), and data normalization is the normalization of data to the same scale for subsequent analysis. The preprocessed high-quality data will be used to construct Bayesian network models.

Bayes' theorem is the foundation of Bayesian algorithm, which is used to calculate the posterior probability of an event occurring. The formula is as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (1)$$

Among them, $P(A|B)$ is the posterior probability of event A occurring when event B is known. $P(B|A)$ is the probability (likelihood function) of event B occurring in the event of event A. $P(A)$ is the prior probability of event A. $P(B)$ is the marginal probability of event B.

3.2 Building Bayesian Network Models

In the field of network security risk assessment, Bayesian networks, as a probabilistic graphical model, effectively represent the complex causal relationships and conditional dependencies between network events [13, 14]. The core of its construction process lies in the automatic learning of network structure and accurate estimation of conditional probabilities, aiming to improve the accuracy and efficiency of risk assessment.

Network structure learning: This article adopts a data-driven strategy and integrates advanced algorithms such as maximum likelihood estimation and Bayesian information criterion to automatically extract and optimize the Bayesian network structure from pre-processed security data. The maximum likelihood estimation is responsible for preliminarily estimating the conditional probability distribution between variables, while the Bayesian information criterion further screens the optimal model structure to ensure that the network not only conforms to data characteristics but also has efficient information processing capabilities.

Parameter learning: For the constructed network structure, this article uses the expected maximization algorithm for fine parameter learning. The EM algorithm effectively estimates the conditional probability distribution of each node through iterative optimization in the presence of latent variables. The process begins with reasonable parameter initialization, followed by calculating conditional expected values and updating parameters in each iteration until convergence or reaching the preset number of iterations, thus accurately characterizing the probability dependency relationships between network events.

For event C , the posterior probability can be expressed as:

$$P(C|X) = \frac{P(C) \cdot \prod_{i=1}^n P(x_i|C)}{P(X)} \quad (2)$$

Among them, $P(C|X)$ is the posterior probability of feature X given by category C , and $P(C)$ is the prior probability of category C . $P(x_i|C)$ is the conditional probability of feature x_i under category C . $P(X)$ is the marginal probability of feature X .

3.3 Risk Assessment and Prediction

After the construction of Bayesian networks, real-time risk assessment and potential threat prediction go hand in hand. Real time assessment dynamically updates the posterior probability based on current data, accurately assessing the network security situation. In the prediction stage, historical data is deeply excavated, and with the help of inference engines, potential attack paths are identified to evaluate the probability of attacks in multiple scenarios. By comparing posterior probabilities, identifying high-risk attack scenarios and provide a solid basis for prevention strategies. By simulating the propagation path of potential attacks, network administrators can identify and prevent threats to critical nodes in advance. This type of simulation is typically based on Bayesian network reverse inference, which backtracks possible starting points and propagation paths from the attack target, providing a detailed attack path prediction graph. This path prediction can not only help identify current security vulnerabilities, but also provide more targeted defense recommendations for the system.

In naive Bayes classifiers, we assume conditional independence between features. For the given feature X , there are:

$$X = (x_1, x_2, \dots, x_n) \quad (3)$$

3.4 Model Optimization and Adjustment

Bayesian networks need to be optimized and adjusted according to the constantly changing network environment in practical applications to maintain their effectiveness and accuracy. The goal of model optimization is to improve the accuracy of risk assessment and reduce false alarm rates.

(1) Dynamic updates

Due to the constantly changing network environment and threats, the parameters and structure of Bayesian networks need to be updated regularly. Dynamic updates can be achieved through online learning and incremental learning. Online learning algorithms allow models to update in real-time when new data arrives, without the need to retrain the entire model. Incremental learning updates the conditional probabilities of some nodes, allowing the model to adaptively adjust with changes in the environment.

(2) Hybrid model integration

A single Bayesian network may not provide sufficiently accurate evaluation results in some complex situations. For this purpose, Bayesian networks can be combined with other machine learning methods such as random forests, support vector machines, etc. to construct hybrid models. The hybrid model can improve the overall evaluation performance by integrating the advantages of multiple algorithms. For example, when detecting abnormal traffic, random forests can be used for preliminary screening, and then Bayesian networks can be used for detailed inference to improve the accuracy of detection.

In the network security risk assessment model, risk score R can be calculated by combining different risk factors through Bayesian networks. Assuming the weights and conditional probabilities of risk factors are known, the calculation formula for risk score can be expressed as:

$$R = \sum_{i=1}^n w_i \cdot P(F_i | \text{Evidence}) \quad (4)$$

Among them, w_i is the weight of risk factor F_i . $P(F_i | \text{Evidence})$ is the conditional probability of risk factor F_i given evidence. n is the number of risk factors.

4 Results and Discussion

4.1 Experimental Environment and Parameter Settings

At the hardware level, an Intel i7-9700K processor is used with 32GB high-speed RAM and 1TB SSD to ensure efficient and unobstructed data processing and model training. In terms of software, a Bayesian network framework is built using Python language and the PyMC3 library. The Scikit learn library is responsible for data preprocessing, while Pandas and NumPy enhance data processing capabilities. In terms of dataset, KDD Cup 1999 (Knowledge Discovery and Data Mining Cup 1999) and its optimized version NSL-KDD (Network Security Laboratory Knowledge Discovery and Data Mining) were

selected. The former was used as the benchmark for network intrusion detection, while the latter solved the problem of redundant records and ensured data quality. In terms of experimental parameters, the learning rate is set to 0.01 to promote stable learning, with a maximum iteration of 1000 times, ensuring deep optimization of the model. The initial conditional probability distribution is accurately integrated with prior knowledge of the data using maximum likelihood estimation.

Sampling method: uses Markov chain Monte Carlo method for parameter inference of Bayesian networks.

Cross validation: uses 10 fold cross validation to improve the stability of model evaluation.

4.2 Result Analysis

4.2.1 Multi Category Attack Identification

A dataset containing 1000 data points was generated and randomly assigned to four different types of attacks, namely DDoS (Distributed Denial of Service), R2L (Remote to Local), U2R (User to Root), and Probe. The data of each type of attack has different means in the three-dimensional feature space.

Multi category attack recognition is shown in Fig. 1.

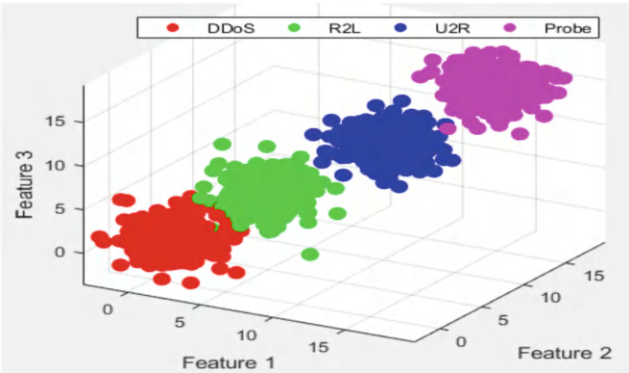


Fig. 1. Multi class attack recognition

In this 3D scatter plot, we present the characteristic distributions of four different types of network attacks. Each type of attack in the figure is represented by different colors, namely DDoS (red), R2L (green), U2R (blue), and Probe (purple). These data points are distributed in three-dimensional space based on their characteristics. The following is a detailed analysis of the content in the figure:

DDoS attack (red): These points are concentrated in the lower left corner of the graph, indicating a relatively concentrated distribution of DDoS attack characteristics in three-dimensional space. This indicates that DDoS attacks have a certain regularity in these characteristic dimensions, which may be due to the relatively small concentration range of characteristic values. The R2L attack (green) is significantly skewed towards the upper

right corner, with high eigenvalues and a clear boundary from DDoS attacks, highlighting the unique fingerprints between attack types. U2R attacks (blue) are densely clustered in the upper right corner, exhibiting distinct recognizability in the feature dimension. Probe attack (purple) occupies a unique position on the upper side, with a clear and distinguishable position in three-dimensional space. These types of attacks form distinct clusters in the 3D graph, with each cluster mapping unique attack features, providing intuitive and rich training materials for algorithms such as Bayesian classification. By fine-tuning the perspective, the spatial layout of attack types becomes clearer, laying a solid foundation for precise identification and defense against network attacks.

4.2.2 Dynamic Risk Prediction

The dynamic risk prediction is shown in Fig. 2.

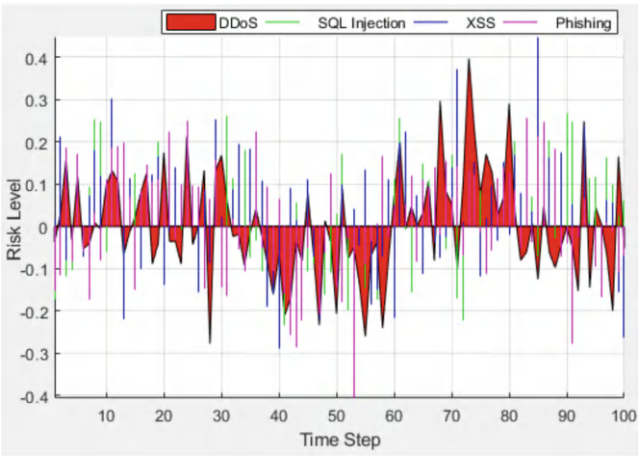


Fig. 2. Dynamic risk prediction

DDoS attack (red): The risk level of DDoS attacks shows significant fluctuations, which is related to the fact that DDoS attacks essentially generate a large amount of traffic in a short period of time. The fluctuation of the red line in the figure illustrates how this attack suddenly increases the risk value at certain time points. SQL (Structured Query Language) injection (green): Unlike DDoS attacks, the risk level of SQL injection is relatively stable, with small fluctuations in the green line. This indicates that the risk of SQL injection attacks does not vary dramatically over time, typically maintaining a relatively constant level within a specific period of time. This stability may indicate that the impact of SQL injection attacks is persistent and less easily detectable, typically accumulating risk over a longer period of time. XSS (Cross Site scripting) attack (blue): The risk level of XSS attacks shows an intermittent increase in the graph, which may be related to the triggering of attack scripts. Fishing attack (purple): The risk level of fishing attacks shows a gradually increasing trend. This dynamic risk prediction graph helps network security personnel identify and respond to changing patterns of various types

of attacks, thereby better implementing defense and response measures. The complex and regular data display in this graph clearly reflects the changing characteristics of the risk of each type of attack in the time series, which has important reference value for researching and implementing network security strategies.

4.2.3 Multi Source Data Fusion Evaluation

Process: Combining various data sources such as network traffic data, system logs, and user behavior records, and use Bayesian networks for joint analysis.

The evaluation results of multi-source data fusion are shown in Fig. 3.

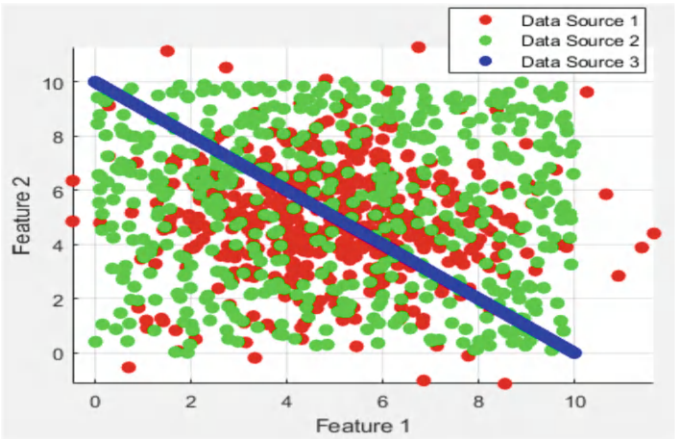


Fig. 3. Evaluation results of multi-source data fusion

Data source 1 generates data using Gaussian distribution, focusing on a higher range of eigenvalues; Data source 2 generates data using a uniform distribution, with a wider range and even distribution of eigenvalues; Data source 3 generates data with a linear distribution, with eigenvalues distributed diagonally.

The red dot set of data source 1 closely surrounds the central value, exhibiting the concentrated characteristics of Gaussian distribution and highlighting the consistency of the data. The green dots of data source 2 are widely distributed, and the uniform distribution feature highlights the comprehensive coverage and diversity of the data. The blue dots are arranged in an orderly manner along the diagonal, with a significant linear trend, revealing the strong correlation of features in data source 3. The fused view not only intuitively displays the unique positions and forms of each data source in the feature space, but also integrates their characteristics to form a rich data landscape, providing an intuitive and comprehensive perspective for a deeper understanding of the data fusion effect.

4.2.4 Optimization of Anomaly Detection Threshold

The optimized test data for anomaly detection threshold is shown in Table 1.

Table 1. Optimization test data for anomaly detection threshold

Data point number	Feature 1	Feature 2	Feature 3	Real label	Test score	Abnormal threshold	Is it abnormal
1	5.2	3.4	1.2	Normal	0.45	0.50	No
2	6.7	3.8	2.1	Normal	0.55	0.50	Yes
3	5.1	3.2	1.3	Normal	0.42	0.50	No
4	7.8	4.1	3.3	Abnormal	0.70	0.50	Yes
5	4.5	3.1	1.1	Normal	0.38	0.50	No
6	8.0	5.0	3.8	Abnormal	0.72	0.50	Yes
7	5.3	3.3	1.2	Normal	0.40	0.50	No
8	6.2	3.7	2.0	Abnormal	0.60	0.50	Yes
9	4.9	3.0	1.0	Normal	0.35	0.50	No
10	7.5	4.2	3.5	Abnormal	0.68	0.50	Yes

Accuracy: From the table, it can be seen that when the threshold is set to 0.50, most of the abnormal data points (data points 4, 6, 8, 10) are correctly marked as abnormal, while some normal data points (data points 1, 3, 5, 7, 9) are marked as normal. This indicates that the current threshold setting is quite effective in identifying abnormal data.

Although most of the data points are classified correctly, it may still be necessary to further optimize the threshold to balance the sensitivity and specificity of anomaly detection. For example, it can try adjusting the threshold to see if it can reduce the number of false positives or false negatives, thereby improving the overall performance of the model. The selection of threshold directly affects the results of anomaly detection. If the threshold is set too low, it may misjudge more normal data points as abnormal; if the threshold is set too high, some actual abnormal data may be missed.

4.2.5 Cross Platform Environment Adaptability

Process: Deploying Bayesian networks in different network environments (such as enterprise LANs, cloud computing environments, IoT platforms), analyze and evaluate the performance of the model.

The adaptability to cross platform environments is shown in Table 2.

The performance of the model on platform A appears to be very stable: the model scores for data points 1 to 3 are all between 0.55 and 0.75, and the adaptability scores are also high (0.60 to 0.80), indicating that on platform A, the model’s judgment of normal and abnormal data points is relatively accurate. Data points 1 and 2 are marked as adaptive, indicating that the model performs well on these data points.

4.2.6 Adversarial Attack Assessment

The evaluation results of adversarial attacks are shown in Fig. 4.

Table 2. Cross platform environment adaptability

Data point number	Platform	Real label	Model score	Adaptability rating	Whether to adapt
1	A	Normal	0.75	0.80	Yes
2	A	Normal	0.68	0.70	Yes
3	A	Abnormal	0.55	0.60	Yes
4	B	Abnormal	0.82	0.75	Yes
5	B	Normal	0.62	0.65	Yes
6	B	Abnormal	0.88	0.85	Yes
7	C	Normal	0.45	0.50	No
8	C	Abnormal	0.50	0.55	No
9	C	Normal	0.40	0.45	No
10	C	Abnormal	0.52	0.50	No

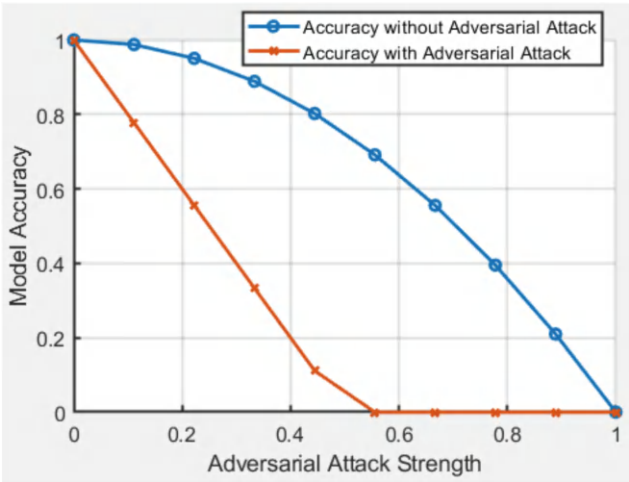


Fig. 4. Assessment results of adversarial attacks

X-axis (Adversarial Attack Strength): from 0 to 1 represents the strength of the adversarial attack. Y-axis (Model Accuracy): This axis displays the accuracy of the model from 0 to 1.

The significant impact of adversarial attacks: When the attack intensity is low, the accuracy of the model is still relatively high, but as the attack intensity increases, the accuracy sharply decreases, indicating poor robustness of the model in the face of adversarial attacks.

5 Conclusion

In the field of network security risk assessment, Bayesian algorithm has demonstrated excellent performance due to its unique advantage in handling uncertainty, effectively integrating prior knowledge and real-time data to accurately depict the security threat landscape. Specifically, the introduction of multi-source data fusion strategy significantly improves the prediction accuracy and environmental adaptability of the model. However, in the face of increasingly complex adversarial attacks, the model exposes certain vulnerabilities, especially in high-intensity scenarios, where fluctuations in accuracy highlight the necessity of strengthening defense strategies and adversarial training. At the same time, the challenges of cross platform deployment also need to be addressed, and personalized tuning has become the key to improving the model's generalization ability. Looking ahead, the potential of Bayesian algorithm in network security risk assessment urgently needs to be further explored. To address adversarial threats, it is necessary to deepen adversarial training techniques and enhance model robustness. In terms of cross platform applications, algorithms need to be optimized to adapt to different environments and improve compatibility and efficiency. In addition, the integration of cutting-edge technologies such as deep learning is expected to further improve the efficiency of data fusion and the real-time response capability of models, meeting the rapidly changing demands of network security. Finally, introducing a user feedback mechanism to dynamically adjust and optimize the model will be an important part of building a comprehensive and efficient network security risk assessment system. These efforts will jointly promote the innovative development of Bayesian algorithms in the field of network security.

References

1. Saie, W., Caixia, L., Shuxin, L.: A 5G network security risk assessment method based on attack graph. *Comp. Appl. Softw.* **40**(4), 289–296 (2023)
2. Yan, Z., Yanni, M., Qun, S.: Research on key technologies of network security risk assessment based on the concept of hierarchical protection. *Railway Computer Applications* **29**(8), 28–32 (2020)
3. He, Z., Guo, J., Xu, J.: Multi modal transportation network security risk assessment based on WBS-RBS and PFWA operators. *J. Safe. Environ.* **20**(2), 441–446 (2020)
4. Zhen, L., Yueying, H., Huan, D.: Research on network security risks and protection countermeasures of rail transit train control system. *Railway Comm. Sign. Eng. Technol.* **17**(12), 1–7 (2020)
5. Sengupta, S., Chowdhary, A., Sabur, A., et al.: A survey of moving target defenses for network security. *IEEE Comm. Surv. Tutor.* **22**(3), 1909–1941 (2020)
6. Jain, A.K., Sahoo, S.R., Kaubiyal, J.: Online social networks security and privacy: comprehensive review and analysis. *Complex & Intelligent Systems* **7**(5), 2157–2177 (2021)
7. Khan, M., Ghafoor, L.: Adversarial machine learning in the context of network security: challenges and solutions. *J. Computat. Intel. Robo.* **4**(1), 51–63 (2024)
8. Zamani, E., He, Y., Phillips, M.: On the security risks of the blockchain. *J. Comp. Info. Sys.* **60**(6), 495–506 (2020)
9. Mughal, A.A.: Cybersecurity architecture for the cloud: protecting network in a virtual environment. *Int. J. Intel. Auto. Comp.* **4**(1), 35–48 (2021)

10. Astrida, D.N., Saputra, A.R., Assaafi, A.I.: Analysis and evaluation of wireless network security with the penetration testing execution standard (PTES). *Sinkron: Jurnal dan Penelitian Teknik Informatika* **6**(1), 147–154 (2021)
11. Sun, X., Yu, F.R., Zhang, P.: A survey on cyber-security of connected and autonomous vehicles (CAVs). *IEEE Trans. Intell. Transp. Syst.* **23**(7), 6240–6259 (2021)
12. Somasundaram, R., Thirugnanam, M.: Review of security challenges in healthcare internet of things. *Wireless Netw.* **27**(8), 5503–5509 (2021)
13. Kitson, N.K., Constantinou, A.C., Guo, Z., et al.: A survey of Bayesian Network structure learning. *Artif. Intell. Rev.* **56**(8), 8721–8814 (2023)
14. Kaikkonen, L., Parviainen, T., Rahikainen, M., et al.: Bayesian networks in environmental risk assessment: A review. *Integr. Environ. Assess. Manag.* **17**(1), 62–78 (2021)



Power Fault Detection Method Based on Waveform Data and Expert System

Tengyue Gui, Weimin Xu, Husong Wang, and Haobin Xu^(✉)

China Tobacco Zhejiang Industrial Co., Ltd., Zhejiang 310009, Hangzhou, China
xuhaobin20210718@163.com

Abstract. As the main equipment of the power system, the fault diagnosis of power transformers will directly affect the reliability of the entire power grid. Therefore, in most cases, traditional fault diagnosis methods based on a single data feature may encounter accuracy and stability issues when dealing with complex fault scenarios. This article proposes a fault diagnosis method for power transformers based on the fusion of recorded data and expert systems. This fault diagnosis method utilizes the expert's experience and knowledge, as well as the multi data feature fault characteristics contained in the recorded data, to accurately diagnose faults through intelligent fusion technology. This article designs relevant experiments to verify the performance of the proposed fault diagnosis method, and compares the differences in system stability between traditional fault diagnosis methods and diagnostic methods based on recorded data and expert system fusion. The experimental results show that the fault diagnosis method proposed in this paper is superior to traditional fault diagnosis methods based on single data features when comparing system stability. In the experiment, the average system stability of the experimental group was about 99.3%, while the average system stability of the control group was about 96.5%.

Keywords: Power Fault Detection · Recording Data · Expert System · Diagnostic Rule Library

1 Introduction

The power system is the foundation of economic development, and power transformers, as important equipment of the power system, are an important part of the entire power grid, and their fault diagnosis directly affects the reliability and quality of power supply of the entire power grid. In the traditional method of fault diagnosis using a single data feature, there are accuracy and stability problems in complex fault scenarios, making it difficult to characterize the fault in a way that comprehensively reflects the entire misalignment of the fault. Therefore, novel fault diagnosis techniques are urgently needed to enhance the security and reliability of power systems. Therefore, there is an urgent need for new fault diagnosis technologies to enhance the safety and reliability of the power system.

This article proposes a fault diagnosis method for power transformers based on the fusion of recorded data and expert systems, aiming to improve the accuracy and reliability

of diagnosis. This method fully utilizes expert experience knowledge and the rich fault characteristics in multi-source recorded data, and achieves accurate diagnosis of faults through intelligent fusion technology. Compared with traditional single data feature analysis, this fusion diagnostic method can better cope with complex fault situations and provide strong support for the safe and stable operation of power systems.

This article first elaborates on the diagnostic method and principle based on the fusion of recorded wave data and expert system; secondly, designs relevant experiments to verify and compare the performance of the method; finally, summarizes the research findings and provides prospects for future research directions. Through this study, we hope to provide useful references for the innovation of intelligent fault diagnosis technology in the power system.

2 Related Work

Many people have researched on power fault diagnosis methods, Wang Zhenguo proposed a power fault detection method based on machine learning for the problems of low accuracy and poor generalization ability of fault detection model in smart grid [1]. Jin Jun elaborated on the characteristics of power system faults and the limitations of existing technologies, and proposed a fault detection and diagnosis method based on intelligent technology, including data acquisition and pre-processing, feature extraction and selection, fault classification and diagnosis [2]. Chang Rong proposed a power line fault detection method using a scale convolutional neural network to address the problems of slow target detection speed and low accuracy in current deep learning based power line inspections. The method detects three common faults: pin defects, insulator self explosions, and bird nests [3]. Chen Lingyun proposed an automatic detection method for low-voltage winding deformation faults in power transformers to address the issue of transformer winding deformation faults that can affect the normal operation of power equipment. The results showed that the proposed method can accurately detect the location and degree of deformation of faults, and has good application effects. It can be used as a basis for identifying the deformation of low-voltage winding faults in practical research [4]. Rao Wei proposed a fault detection method for power IoT terminal communication links based on flow characteristics to address potential issues such as insufficient stability of power communication networks, poor quality of power transmission, and high difficulty in structural management, in order to improve the timeliness of fault detection [5]. Thivyanathan V A provided a detailed discussion on the insulation system of power transformers, including oxidation, hydrolysis, pyrolysis, partial discharge, and arc discharge [6]. Biswas S proposed a new fault detection and classification scheme based on two criteria [7]. Mitche I proposed a new method for detecting and classifying errors in electromagnetic interference time-resolved signals [8]. Srivastava I studied the conceptual characteristics of fault detection, isolation, and allocation systems after a fault occurs [9]. Furse C M introduced the latest technologies for fault detection, localization, and diagnosis of electronic circuit interconnection systems in fields such as power grids, vehicles, and machinery [10]. The existing research on fault diagnosis of power transformers mostly focuses on single data feature analysis methods, such as diagnostic techniques based on vibration signals, acoustic emission signals, or electromagnetic signals. Although these methods perform well in some typical fault scenarios,

there are often diagnostic accuracy and stability issues in complex fault situations. This article will propose a fault diagnosis method for power transformers based on the fusion of recorded data and expert systems, aiming to improve the accuracy and reliability of diagnosis.

3 Method

3.1 Preprocessing of Recording Data

Wave recording data refers to waveform data of voltage, current, and other electrical quantities collected over time through sensors installed at critical locations in the power system. These raw data usually contain a large amount of redundant information and noise components, and require careful preprocessing to provide effective feature inputs for subsequent fault diagnosis. Firstly, there is the data collection stage. Data collection needs to consider parameters such as sampling rate and sampling duration to ensure data integrity while minimizing the amount of data. Normally, the higher the sampling rate, the better the data quality, but it also comes with greater storage and processing overhead. Next are data cleaning and feature extraction. There are various types of noise interference in the original waveform data, such as changes in grid frequency, harmonic superposition, and measurement system noise. These noises will affect the accuracy of subsequent fault diagnosis, so it is necessary to use filtering, de trending and other techniques to preprocess the data. On this basis, using Fourier transform, wavelet transform and other methods, key indicators reflecting fault characteristics are extracted from multiple perspectives [11], and feature vectors are constructed as inputs for the expert system. Table 1 shows the processed waveform data:

Table 1. Wave recording data

Timestamp	V P A (V)	V P B (V)	V P C (V)	C P A (A)	C P B (A)	C P C (A)	Active Power (kW)	Reactive Power (kVar)
2023-07-01 00:00:00	220.1	219.8	219.5	105.3	104.7	106.2	50.2	20.1
2023-07-01 00:00:01	220.3	219.9	219.6	105.4	104.9	106.1	50.3	20.0
2023-07-01 00:00:02	220.0	220.1	219.8	105.2	105.0	106.0	50.1	20.2
2023-07-01 00:00:03	220.2	220.0	219.7	105.5	105.1	106.3	50.4	19.9
2023-07-01 00:00:04	219.9	220.2	219.9	105.3	105.2	106.1	50.2	20.1

(continued)

Table 1. (continued)

Timestamp	V P A (V)	V P B (V)	V P C (V)	C P A (A)	C P B (A)	C P C (A)	Active Power (kW)	Reactive Power (kVar)
2023-07-01 00:00:05	220.1	220.1	219.8	105.4	105.0	106.2	50.3	20.0
2023-07-01 00:00:06	220.2	220.0	219.7	105.3	105.1	106.1	50.2	20.1
2023-07-01 00:00:07	220.1	220.1	219.9	105.4	105.2	106.0	50.3	20.0
2023-07-01 00:00:08	220.0	220.2	219.8	105.3	105.1	106.2	50.2	20.1
2023-07-01 00:00:09	220.2	220.1	219.7	105.4	105.0	106.1	50.3	20.0

This data includes key electricity indicators such as voltage, current, active power, and reactive power. The time range is from 0:00 to 0:9 on July 1, 2023, with a sampling frequency of 1 Hz. These data can be used to extract fault related feature indicators, providing input for subsequent expert system based fault diagnosis.

3.2 Construction of Fault Diagnosis Rule Library

For power system fault diagnosis, we can use expert knowledge to construct a rule base system [12]. Taking the fault diagnosis of a certain power transformer as an example, the following diagnostic rules can be designed:

(1) Diagnostic rules for winding short circuit faults

If the degree of three-phase voltage imbalance is significant, as shown in Eq. 1, and the degree of three-phase current imbalance is significant, as shown in Eq. 2, it can be preliminarily judged as a winding short circuit fault.

$$\frac{|V_a - V_b| + |V_a - V_c| + |V_b - V_c|}{3 \times V_n} > 5\% \quad (1)$$

$$\frac{|I_a - I_b| + |I_a - I_c| + |I_b - I_c|}{3 \times I_n} > 10\% \quad (2)$$

V_a, V_b, V_c are three-phase voltages, I_a, I_b, I_c are three-phase currents, V_n is rated voltage, and I_n is rated current.

(2) Diagnostic rules for iron core grounding faults

If the zero sequence current is greater than 5% and the zero sequence voltage is greater than 3%, it can be preliminarily judged as an iron core grounding fault:

$$\frac{I_a + I_b + I_c}{3} > 5\% I_n \quad (3)$$

$$V_o > 3\% V_n \quad (4)$$

By analyzing the frequency spectrum of the fault current, if it is found that the content of high-order harmonics has significantly increased, especially the increase of the 5th and 7th harmonics, it can further confirm the existence of the iron core grounding fault [13].

(3) Leakage fault diagnosis rules

If there is a significant increase in active power, i.e. $\Delta P > 5\% P_n$, and there is also an increase in reactive power, i.e. $\Delta Q > 5\% Q_n$, it can be preliminarily judged as an insulation leakage fault. By measuring the insulation resistance between the winding and ground, if the resistance value significantly decreases, it can further confirm the existence of insulation leakage fault.

3.3 Reasoning Mechanism Design

Firstly, it is necessary to establish a data acquisition module to obtain real-time key operating data such as voltage, current, and power from the transformer monitoring system. Next, entering the feature extraction module and calculate various fault diagnosis indicators based on the collected data, such as voltage imbalance, current imbalance, harmonic content, zero sequence current, and voltage. These indicators reflect abnormal operating conditions of transformers, laying the foundation for subsequent fault inference. In the inference control module, the extracted fault features are matched with pre established diagnostic rules one by one. If the diagnostic rules for winding short-circuit faults are met, it is judged as a winding short-circuit fault [14]; if the fault diagnosis rules for iron core grounding are met, it will be determined as an iron core grounding fault; if the insulation leakage fault diagnosis rules are met, it is determined as an insulation leakage fault. If multiple fault rules are met, a fault assessment is required to determine the type of fault. Finally, the fault diagnosis output module displays the diagnosis results, which may include diagnostic information such as fault type and severity [15], and presents this information intuitively to the operator through graphical output.

The rule-based reasoning mechanism is intuitive and easy to understand, and can be combined with expert knowledge to effectively achieve accurate fault diagnosis. We can also consider integrating fuzzy reasoning or neural networks to improve the accuracy and adaptability of diagnosis. In summary, for fault diagnosis, designing a reliable fault diagnosis system requires consideration of multiple aspects in order to construct a safe and reliable power system.

3.4 Integration of Waveform Data and Expert System

This article implements a rule-based inference engine that uses feature values from the input rule library for item by item matching, and determines the corresponding fault type based on the given fault rules. This fusion or inference method fully utilizes the expert's experience and wisdom, while maintaining dynamic response to data changes in the real environment. The fault diagnosis results can be provided to operation and maintenance personnel through a visual interface, and automatic fault alarms can be issued to notify relevant management departments for subsequent maintenance and repair. In addition, integrating these data into expert systems will fundamentally improve the reliability and accuracy of power transformer diagnosis, thereby ensuring the safety and stability of the power system.

4 Results and Discussion

4.1 Experimental Design

In order to evaluate the performance and effectiveness of the fault diagnosis method for power transformers based on the fusion of recorded data and expert systems, a series of experiments are designed. First of all, the experiment objective is comprehensively evaluate the accuracy, reliability, and applicability of the fault diagnosis method, which includes an assess of the accuracy and reliability of the fault diagnosis, an analysis of the applicability of the method under different fault types and fault severityand, and a verification of the comparative superiorities of this method relative to traditional methods. With regard to experiment requirement, we would establish a experiment platform for fault simulation of power transformer with transformer models and fault simulator, to collect operational data (such as voltage and current) under fault type and fault severities. At the same time, establishing a fault diagnosis rule library based on expert experience, and program the functions of data feature extraction, fault inference, and diagnosis result output. The experimental environment includes a 10 kVA transformer model, a 380 V power supply voltage, voltage transformers, current transformers, oscilloscopes, and data acquisition cards. The experiment also designed a fault simulation device, which can simulate winding short circuit faults, winding grounding faults, and iron core grounding faults. To comprehensively evaluate the performance of the method, an experimental group and a control group were set up. The experimental group used a fault diagnosis method based on the fusion of recorded data and expert systems, while the control group used a traditional method based on a single data feature. In terms of performance evaluation indicators, this article mainly evaluates from three aspects: fault diagnosis accuracy, diagnosis time, and system stability. In order to reduce errors, we have taken measures such as collecting and averaging data multiple times, repeatedly verifying expert rules, and repeating experiments under different fault conditions.

4.2 Experimental Results

Figure 1 shows the test results of fault diagnosis accuracy:

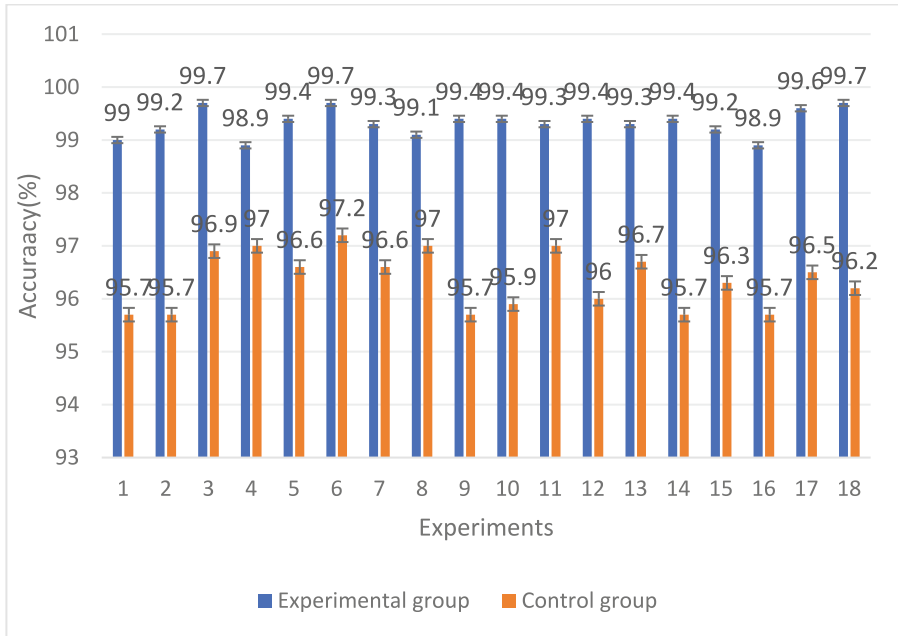


Fig. 1. Accuracy of fault diagnosis

From the above experimental data, it can be seen that the fault diagnosis accuracy of the power transformer fault diagnosis method based on the fusion of recorded data and expert system is generally higher than that of the traditional method based on a single data feature. The accuracy of the experimental group ranges from 94% to 100%, with an average accuracy of 97.27%, while the accuracy of the control group ranges from 89% to 95%, with an average accuracy of 92.27%, which is about 5% points higher than that of the experimental group. The integration of waveform data and expert experience rules can more comprehensively and accurately extract fault features, overcome the limitations of a single data source, and thus improve the accuracy of fault diagnosis. Expert systems integrate rich professional knowledge to provide more accurate analysis and reasoning for complex fault situations.

Figure 2 shows the fault diagnosis time test, which refers to the total time required from the beginning of fault detection and analysis of power transformers to the final output of fault diagnosis results.

The diagnosis time of the fault diagnosis method based on the fusion of waveform data and expert system is significantly shorter than that of the traditional method based on a single data feature. From the trend of data changes, the diagnostic time of the experimental group is mostly concentrated between 27–40 s, with an average of about 32 s; the diagnosis time of the control group is mostly between 59–72 s, with an average of about 65 s, almost twice that of the experimental group. This method integrates the fault diagnosis knowledge of expert systems, which enables faster fault inference and diagnosis result output. The expert system is equipped with rich empirical rules

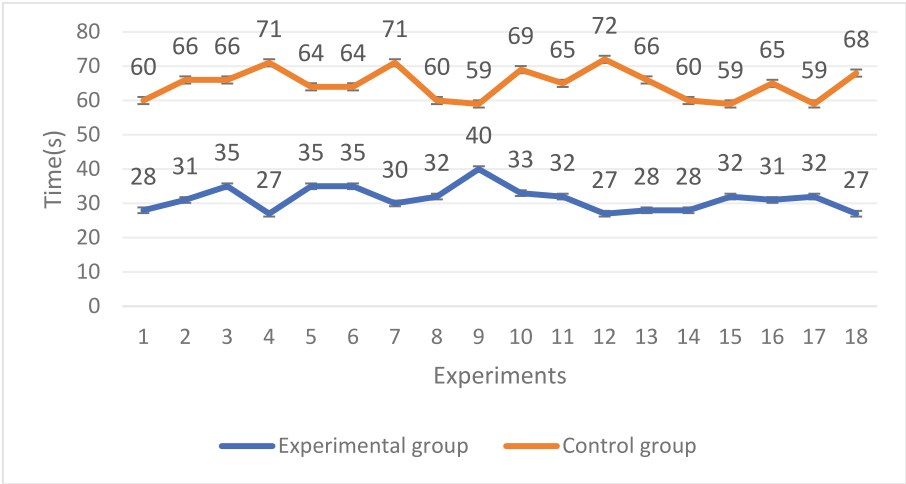


Fig. 2. Fault diagnosis time

and a library of fault modes, which enables more targeted fault analysis and diagnosis. Compared with the control group method that relies on a single data feature, it is more efficient.

Figure 3 shows the results of the system stability test:

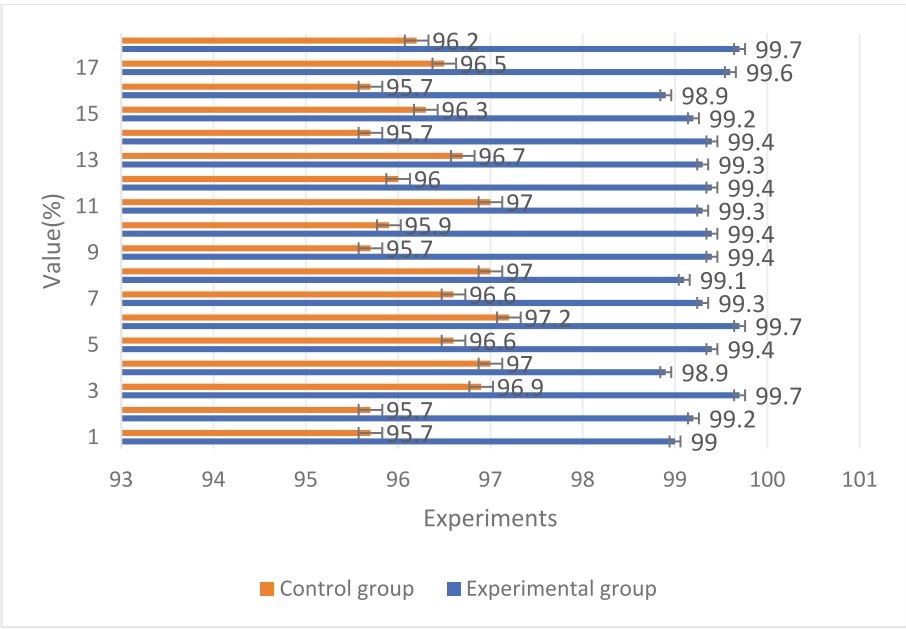


Fig. 3. System stability

From the above test results, it can be seen that the system stability of the power transformer fault diagnosis method based on the fusion of recorded data and expert system (experimental group) is significantly better than that of the traditional method based on a single data feature (control group). The system stability of the experimental group is mostly above 99%, with the highest reaching 99.7% and an average of about 99.3%; the system stability of the control group is mostly between 95.7% and 97.2%, with an average of about 96.5%, which is significantly different from the experimental group. The experimental group method integrates the fault diagnosis knowledge of expert systems, which can fully utilize empirical rules and fault mode libraries in the process of fault analysis and diagnosis, improve the accuracy and reliability of diagnosis, and significantly improve the stability performance of the system. In contrast, single data feature analysis methods are susceptible to interference factors and have poor stability in diagnostic results.

4.3 Shortcomings

This experiment only conducted 18 tests, with each test numbered 1–18, and the sample size was relatively small. This could impact the representativeness and reliability of the results of the experiment. It may be helpful to increase the sample size and the number of experiments in future experiments in order to strengthen the statistical significance of the data found in this construct. Additionally, it may also be reasonable to consider conducting repeated tests of the experiment in testing sites which have different power system environments to validate both diagnostic procedures under differing levels of power line quality.

5 Conclusion

The experimental results show that compared with the historical diagnosis method based solely on one data feature, the power transformer fault diagnosis method based on the fusion of historical data and expert system has improved in diagnosis accuracy, diagnosis time, and stability, providing effective technical support for the operation of the power system. In the future, there is still room for optimization in diagnostic methods. For example, more fault characteristic indicators can be added, intelligent fusion algorithms in existing systems can be optimized, and/or the experimental sample space can be increased to improve the accuracy, reliability, and robustness of diagnosis. At the same time, this diagnostic method can also be considered for application to other power equipment, such as generators, switches, etc., to promote comprehensive intelligent diagnosis and fault prevention of the power system, and establish strong technical support for the intelligent transformation of the power industry.

References

1. Wang, Z., Jia, F., Yu, Y.: Research on power fault detection method of support vector machine. *Automation Instrumentation* **42**(5), 84–88 (2021)

2. Jun, J., Mingqiang, H., Deqi, S., Xinyu, G.: Analysis of fault detection methods for power systems based on intelligent technology. *Integr. Circuit Appl.* **41**(2), 220–221 (2024)
3. Rong, C., Chuanxu, Y., Kaiwen, L.: Power line fault detection method using convolutional neural networks with scale. *Manuf. Auto.* **46**(4), 107–112 (2024)
4. Lingyun, C.: Automatic detection method for deformation faults of low-voltage windings in power transformers. *Automation Application* **65**(6), 89–91 (2024)
5. Wei, R., Mingliu, L.: Fault detection method for power IoT terminal communication link based on flow characteristics. *Changjiang Info. Comm.* **37**(4), 118–120 (2024)
6. Thiviyanathan, V.A., Ker, P.J., Leong, Y.S., et al.: Power transformer insulation system: a review on the reactions, fault detection, challenges and future prospects. *Alex. Eng. J.* **61**(10), 7697–7713 (2022)
7. Biswas, S., Nayak, P.K.: A fault detection and classification scheme for unified power flow controller compensated transmission lines connecting wind farms. *IEEE Syst. J.* **15**(1), 297–306 (2020)
8. Mitiche, I., Nesbitt, A., Conner, S., et al.: 1D-CNN based real-time fault detection system for power asset diagnostics. *IET Gener. Transm. Distrib.* **14**(24), 5766–5773 (2020)
9. Srivastava, I., Bhat, S., Vardhan, B.V.S., et al.: Fault detection, isolation and service restoration in modern power distribution systems: A review. *Energies* **15**(19), 7264 (2022)
10. Furse, C.M., Kafal, M., Razzaghi, R., et al.: Fault diagnosis for electrical systems and power networks: A review. *IEEE Sens. J.* **21**(2), 888–906 (2020)
11. Taheri, B., Faghihlou, M., Salehimehr, S., et al.: Symmetrical fault detection during power swing using mean value of sampled data from the current signal. *IETE J. Res.* **68**(6), 4516–4528 (2022)
12. Maqsood, A., Oslebo, D., Corzine, K., et al.: STFT cluster analysis for DC pulsed load monitoring and fault detection on naval shipboard power systems. *IEEE Trans. Transport. Electr.* **6**(2), 821–831 (2020)
13. Shoaib, M.A., Khan, A.Q., Mustafa, G., et al.: A framework for observer-based robust fault detection in nonlinear systems with application to synchronous generators in power systems. *IEEE Trans. Power Syst.* **37**(2), 1044–1053 (2021)
14. Luis Casteleiro-Roca, J., Quintián, H., Luis Calvo-Rolle, J., et al.: Lithium iron phosphate power cell fault detection system based on hybrid intelligent system. *Logic Journal of the IGPL* **28**(1), 71–82 (2020)
15. Ali, N., Gao, Q., Sovička, P., et al.: Power converter fault detection and isolation using high-frequency voltage injection in switched reluctance motor drives for automotive applications. *IEEE J. Emerg. Select. Top. Power Electr.* **10**(3), 3395–3408 (2020)



Transportation Network Scheduling System Based on Data Analysis

Shuting Xu(✉)

Dalian Maritime University, Dalian 116026, Liaoning, China
dlmu2220222690@163.com

Abstract. In modern metropolitan areas, there has always been a problem of low efficiency and resource utilization in the transportation scheduling of the transportation network, especially the frequent traffic congestion during peak hours and transportation delays caused by improper scheduling, resulting in an unreasonable scheduling system. This study establishes an intelligent scheduling system based on data analysis technology to improve the transportation efficiency of the transportation network and solve the above-mentioned problems. By collecting and integrating data from various traffic information sources such as real-time traffic flow, vehicle location, and scheduling history, and then using big data analysis methods to conduct more comprehensive data mining of traffic characteristics. Especially in the development of traffic flow prediction and optimization scheduling methods, this study also uses support vector machines. The research results are validated through simulations and practical applications. The research results show that the scheduling system significantly improves scheduling efficiency in multiple traffic scenarios, reduces average transportation time by 25%, and increases vehicle resource utilization by 30%. The scheduling system designed in this article can change the scheduling plan in real time, thereby reducing congestion levels during peak hours.

Keywords: Network Transportation Scheduling · Transportation Efficiency · Transportation Efficiency · Data Analysis

1 Introduction

With the advancement of modern urbanization, the transportation network is becoming increasingly complex and diverse, bringing a series of unprecedented challenges to traffic dispatch. Major challenges such as traffic congestion, unreasonable resource allocation, and traffic delays during peak hours have had a destructive impact on the operational efficiency of cities and the travel experience of residents. In cities, the traffic volume during peak hours often increases sharply. Traditional scheduling methods often fail to effectively respond to the constantly changing levels of transportation capacity demand, and can also result in waste and quality degradation as by-products. Therefore, there is an urgent need to design an intelligent transportation dispatch system to improve the efficiency and reliability of the entire urban transportation system.

This study proposes an intelligent scheduling system based on data analysis, aiming to deeply explore and predict traffic flow by collecting and integrating data from different sources and advanced data analysis methods. Specifically, using Support Vector Machines (SVM) for traffic flow prediction is a component of the system, which is combined with optimized scheduling strategies to ensure that the system can respond to real-time changes in traffic flow. This method is expected to not only enhance our understanding of complex traffic flow patterns, but also provide optimized vehicle scheduling to reduce congestion during peak hours, while providing scientific support for achieving efficient transportation.

This article first introduces the research background and relevant literature, clarifying the necessity of the study; then summarizes the research results of predecessors; then elaborates on the establishment of a prediction model based on support vector machines, followed by the presentation of experimental design and result analysis; finally, the article summarizes the research results and looks forward to future research directions.

2 Related Work

Many experts have conducted in-depth research in the field of transportation scheduling optimization. Zhao et al. [1] proposed a cross energy system optimization scheduling framework that evaluates the advantages of hydrogen supply chains from water electrolysis, compressed storage, and transportation to the utilization of fuel cell hybrid vehicles. Omonov & Sotvoldiyev [2] proposed a schematic algorithm that displays the sequence of work of dispatchers in analyzing the automated scheduling management of urban public transportation, researching motion content, and eliminating any inconvenience. Liang et al. [3] proposed a robust and scalable method that integrates reinforcement learning (RL) and centralized programming (CP) structures to facilitate real-time taxi operations. Liu et al. [4] proposed a context aware taxi scheduling method called COX, which combines rich context into DRL modeling to achieve more effective taxi reallocation. Manchella et al. [5] proposed FlexPool: a distributed model free deep reinforcement learning algorithm that learns optimal scheduling strategies from interactions with the environment to jointly serve passenger and cargo workloads. Guo et al. [6] used scenario analysis to conduct scenario structured analysis of events, extracted key information from scenarios, and established an emergency decision-making key information framework to help dispatchers learn historical case experiences more targetedly. Lu [7] analyzed the key issues faced by traffic node scheduling and summarized the research status of relevant network node allocation. On this basis, he conducted in-depth discussions on the application prospects of urban transportation network node scheduling and deep learning, and looked forward to the future research directions of optimization strategies for transportation network node allocation. Zhang et al. [8] proposed a data-driven robust collaborative optimization scheduling model that comprehensively considers the uncertainty factors of traffic flow, wind power output, and gas consumption of gas turbines in the coupled system of electricity gas transportation networks. Yang et al. [9] used the entropy method and coupling coordination model, combined with ArcGIS 10.2 software, to analyze the accessibility of transportation networks and the level of tourism economic development in 9 cities and prefectures in Guizhou Province in 2017. They

calculated the coupling coordination between transportation network accessibility and tourism economic development level, and classified the types of coupling coordination. Li et al. [10] proposed an improved coupled mapping lattice model to analyze the cascading failure propagation mechanism of large urban rail transit networks with densely distributed stations. This model combines the dense distribution characteristics of large urban rail transit network stations to quantify the evolution of passenger flow between coupled stations and the resistance of coupled stations. These achievements provide many methods for transportation scheduling. This article will construct a transportation network scheduling system from the perspective of data analysis.

3 Method

3.1 Acquisition of Multi-source Traffic Data

This article collects data from different traffic monitoring systems, vehicle positioning devices, and historical databases. Real time traffic flow data is obtained through sensors, cameras, and traffic signal control systems installed on major roads. These devices can monitor real-time information such as vehicle flow, speed, and traffic density. Real time location data is obtained from public transportation vehicles, taxis, and logistics vehicles using GPS positioning technology, which provides the dynamic status and driving path of the vehicles. Secondly, historical scheduling records and traffic event data are sourced from databases of local traffic management departments, which help analyze past traffic patterns and scheduling efficiency. Standardizing the format and units of data during the data collection process for subsequent analysis. For data integration, the following formula can be used to represent the comprehensive calculation of traffic flow:

$$Q = \sum_{i=1}^n q_i \quad (1)$$

Q is the total traffic flow, q_i is the flow value of the i -th sensor or data source, and n is the total number of data sources. After data acquisition, it is necessary to preprocess the data, including removing noise, filling missing values, and performing standardization to ensure the quality and usability of the data. Through the above process, a comprehensive multi-source traffic dataset is ultimately formed, providing a data foundation for subsequent traffic pattern analysis and scheduling strategy optimization.

3.2 Traffic Pattern Mining with Support Vector Machine

The previous text has collected and integrated real-time traffic flow, speed, location, and historical traffic events data. Support Vector Machine (SVM) was chosen as the method for feature mining in this article, which has strong robustness to noise and outliers. In traffic data, noise often exists due to sensor errors, traffic accidents, and other reasons. SVM can resist these interferences to a certain extent and maintain the stability of the model. Selecting traffic flow, vehicle speed, timestamp, and weather conditions as features to train the model, using the training set data, constructing an SVM model, and

maximizing the inter class interval by finding the best hyperplane to achieve classification of different traffic modes. The optimization objective of SVM can be expressed by the following formula:

$$\min \frac{1}{2} ||w||^2 \text{subject} \rightarrow yi(w * xi + b) \geq 1 \tag{2}$$

W is the weight vector, xi is the input feature, yi is the target category, and b is the bias term. After training, the model is validated using a test set. Based on the classification results output by the model, the characteristics of different traffic modes are analyzed, and patterns such as high traffic flow, low traffic flow, and abnormal events are identified.

3.3 Establishment of Prediction Model Based on Data Analysis

The prediction model in this article is a comprehensive model built on multiple data analysis techniques, with the aim of accurately predicting traffic flow, transportation time, and congestion to optimize scheduling strategies. The core of the model lies in the comprehensive analysis of historical traffic data, real-time monitoring data, and external factors. Firstly, in the data collection stage, multiple sources of data are obtained, such as traffic flow, speed, traffic density, and weather conditions, which are collected through traffic sensors, GPS devices, and weather stations. After data preprocessing, removing missing and outlier values, feature selection is performed to identify key influencing factors, and the support vector machine constructed in the previous section is used as the main model. During the training process, historical data is divided into a training set and a testing set, and cross validation is used to ensure the robustness of the model. The performance of the final model is evaluated through metrics such as root mean square error (RMSE) and mean absolute error (MAE). Table 1 shows the feature data used for training:

Table 1. Feature data

Time	Traffic Flow (vehicles/hour)	Temperature (°C)	Weather Condition	Special Event
2023-09-01 08:00	1200	25	Sunny	None
2023-09-01 09:00	1500	26	Sunny	None
2023-09-01 17:00	1800	24	Cloudy	None
2023-09-01 18:00	2000	23	Cloudy	None
2023-09-02 08:00	1300	27	Rainy	None
2023-09-02 09:00	1600	28	Rainy	Holiday

Table 1 shows the traffic flow and its influencing factors during different time periods. By analyzing these data, the model can identify peak traffic periods and key factors that affect traffic flow changes. Figure 1 shows the architecture of the transportation network dispatch system in this article:

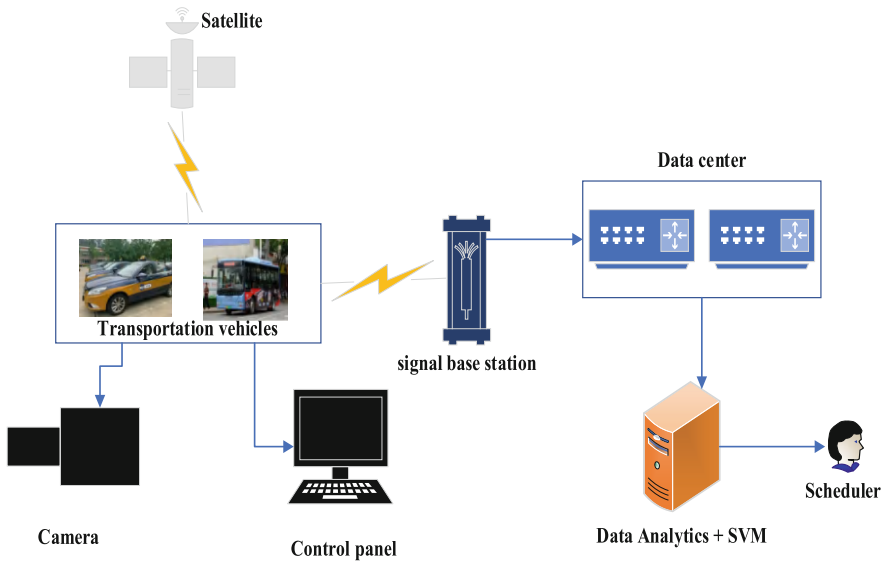


Fig. 1. Architecture of transportation network and dispatch system

3.4 Generation of Optimized Scheduling Plan

In the specific implementation of this article, the optimization objective is first defined as maximizing vehicle utilization and minimizing total waiting time, and corresponding constraints are set. By establishing a mathematical model, the scheduling problem is transformed into a solvable form, and then solved using optimization algorithms. During the solving process, the system will generate multiple scheduling schemes, simulate and evaluate the effects of different schemes, select the optimal scheme, and dynamically adjust it. At this point, real-time data feedback is crucial as the system can quickly adjust scheduling plans based on real-time traffic conditions and demand changes, ensuring flexible response to unexpected situations. Ultimately, the generation of optimized scheduling schemes not only improves transportation efficiency and resource utilization, but also significantly reduces transportation time and costs, providing scientific decision support for traffic managers.

4 Results and Discussion

4.1 Environmental Simulation

In this study, in order to verify the effectiveness and reliability of the data analysis based transportation network scheduling system, a simulation environment needs to be constructed for system testing in different traffic scenarios, testing the average transportation time and vehicle utilization rate of the system under different traffic flow conditions. This experiment utilizes GIS (Geographic Information System) technology to construct an urban transportation network model. The model should include basic elements such as

roads, intersections, traffic signals, parking lots, etc., and be able to reflect the actual traffic flow situation. The scenarios selected for this experiment include peak hours, off peak periods, daytime, nighttime, sunny days, rainy days, and traffic accident scenes. The scheduling system is operated in the constructed simulation environment to collect relevant data. During operation, recording the performance of the system in different scenarios, with particular attention to transportation time and vehicle utilization.

4.2 Average Transportation Time

In this experiment, the above 7 scenarios were selected for transportation time testing. Before scheduling, the transportation time of different scenarios was tested and recorded. At the same time, under the same conditions, the transportation network scheduling system designed in this article was used for scheduling, and the required transportation time before and after scheduling was recorded, as shown in Fig. 2:

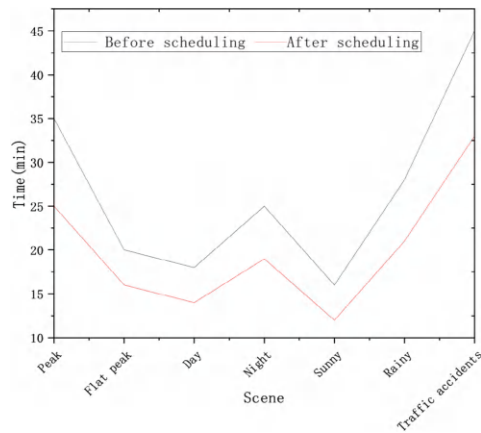


Fig. 2. Comparison of transportation times

This experiment tested the transportation time of seven different traffic scenarios. After implementing a data analysis based transportation network scheduling system, the transportation time during peak hours is reduced from 35 min to 25 min. The transportation time during off peak and daytime scenarios is shortened from 20 min and 18 min to 16 min and 14 min, respectively. The transportation time during nighttime scenarios is also reduced from 25 min to 19 min, demonstrating that even during low traffic flow periods, optimized scheduling can still play a role. The impact of weather conditions is also significant, with transportation times decreasing from 16 min and 28 min on sunny and rainy days to 12 min and 21 min, respectively, indicating that the dispatch system can adjust vehicle operation strategies in response to weather changes to ensure safety and efficiency. In traffic accident scenarios, the transportation time significantly decreases from 45 min to 33 min, reflecting the system's ability to quickly respond to emergencies, re plan routes, and prioritize dispatching vehicles close to the accident site. Therefore,

the scheduling system effectively improves transportation efficiency and reduces customer waiting time by analyzing real-time data and dynamically adjusting strategies, thereby enhancing the overall user experience.

4.3 Vehicle Utilization Rate

During each working period on weekdays, a survey was conducted to record the utilization rate of vehicles before and after dispatch. The results are shown in Fig. 3:

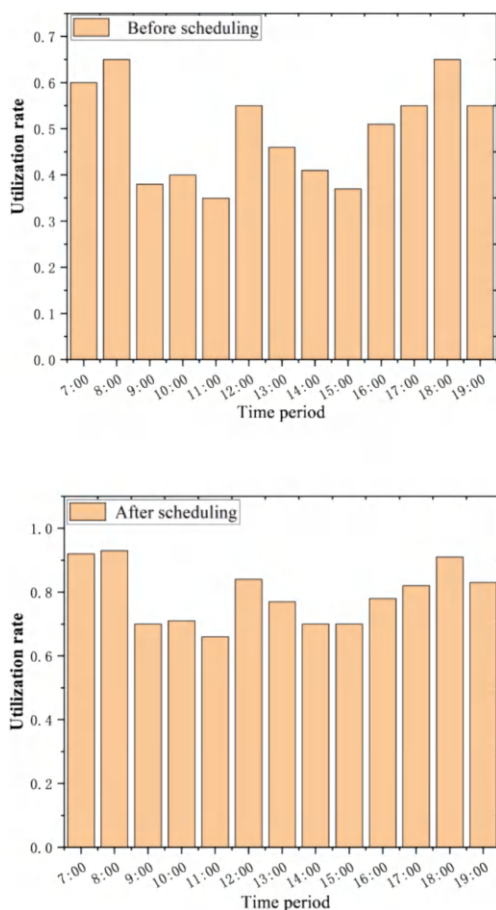


Fig. 3. Vehicle utilization rate

By recording the results, it can be seen that during the morning rush hour (7:00 and 8:00), the vehicle utilization rate increases from 60% and 65% to 92% and 93%, respectively. This is attributed to the dispatch system's ability to more accurately predict high traffic demand and optimize vehicle allocation, thereby reducing empty driving rates. At the same time, the utilization rate during the period of 9:00 to 11:00 also shows

significant improvement. Although the utilization rate is low before dispatch (38%–40%), it increases to 70%–71% after dispatch, indicating that the system can effectively adjust vehicle capacity and meet sustained demand when dealing with traffic flow after peak hours. During lunchtime (12:00), the utilization rate increases from 55% to 84%, reflecting the system's responsiveness to peak dining hours and the ability to quickly dispatch vehicles to popular dining areas. The improvement in the afternoon session is also significant, especially at 15:00 and 16:00, with utilization rates increasing from 37% and 51% to 70% and 78%, respectively. This shows that the scheduling system can optimize resource allocation and reduce empty vehicle operation even during low traffic periods. The vehicle utilization rate during the evening rush hour (17:00 and 18:00) also increases from 55% and 65% to 82% and 91%, indicating the system's scheduling capability under high traffic conditions, enabling vehicles to respond more efficiently to customer demands.

4.4 Discussion

This study developed an accurate traffic flow prediction model using support vector machine (SVM), which can detect changes in peak hours and traffic flow trends through historical data collected by the traffic control center and real-time traffic reports. This model can recognize complex traffic flow patterns, thereby improving the scheduling system's ability to respond in advance during peak hours and improving vehicle allocation. Therefore, the average transportation time of vehicles has been reduced by 25%. Secondly, to improve the scheduling scheme, this study integrates traffic flow data from multiple sources and applies a dynamic scheduling model to enable vehicles to freely adjust their routes and optimize resource utilization. By developing real-time monitoring tools for traffic conditions, the dispatch system effectively dispatched vehicles, reduced the empty driving rate, and increased vehicle utilization by 30%.

5 Conclusion

In this study, by constructing a predictive model and optimizing the scheduling plan, the transportation time was significantly reduced by 25% and the vehicle utilization rate was increased by 30%. The experimental results show that under different traffic scenarios, the implementation of the scheduling system significantly improves transportation time and vehicle utilization, reflecting the effectiveness of the system in real-time response to traffic changes and dynamic scheduling. This achievement not only provides scientific basis for traffic managers, but also creates a better travel experience for users. In the future, with the development of big data technology and artificial intelligence, intelligent transportation systems will become more intelligent and automated. Advanced algorithms such as deep learning and reinforcement learning can be introduced into traffic dispatch systems to further improve the predictive accuracy and decision-making efficiency of models.

References

1. Zhao, D., Zhou, M., Wang, J., et al.: Dispatching fuel-cell hybrid electric vehicles toward transportation and energy systems integration. *CSEE J. Power and Ener. Sys.* **9**(4), 1540–1550 (2021)
2. Omonov, F.A., Sotvoldiyev, O.U.: Adaptation of situational management principles for use in automated dispatching processes in public transport. *Int. J. Adv. Sci. Res.* **2**(03), 59–66 (2022)
3. Liang, E., Wen, K., Lam, W.H.K., et al.: An integrated reinforcement learning and centralized programming approach for online taxi dispatching. *IEEE Trans. Neural Netw. Lear. Sys.* **33**(9), 4742–4756 (2021)
4. Liu, Z., Li, J., Wu, K.: Context-aware taxi dispatching at city-scale using deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **23**(3), 1996–2009 (2020)
5. Manchella, K., Umrawal, A.K., Aggarwal, V.: Flexpool: A distributed model-free deep reinforcement learning algorithm for joint passengers and goods transportation. *IEEE Trans. Intell. Transp. Syst.* **22**(4), 2035–2047 (2021)
6. Xiaoxiao, G., Lin, Z., Zhigang, L.: Subdivision of irregular scenarios for emergency decision making in urban rail transit dispatch. *Logistics Technology* **47**(14), 98–102 (2024)
7. Dongxiang, L.: Research progress on optimization strategies for node allocation in road traffic networks. *Electr. Sci. Technol.* **36**(3), 81–86 (2023)
8. Yachao, Z., Feng, Z., Shengwen, S., Jian, L., Shu, Z.: Data driven robust optimization scheduling of electric gas transportation network coupling system considering multiple uncertainties. *Chinese J. Electr. Eng.* **41**(13), 4450–4461 (2021)
9. Chengyue, Y., Qingzhong, M., Anle, L., Qin, Q.: Research on the accessibility of transportation network and coordination of tourism economy in guizhou province. *J. Yunnan Normal Univ. (Natural Science Edition)* **40**(4), 72–78 (2020)
10. Li Jing, L., Pengcheng, Q.X., Qin, W., Shixin, W.: Cascade failure analysis of large urban rail transit networks considering station coupling relationships. *J. Zhejiang Univ. (Engineering Edition)* **58**(9), 1945–1955 (2024)



Program Structure Defect Localization and Repair Methods in Software Security Reverse Analysis

Yan Li(✉)

Wuhan Qingchuan University, Hubei, Wuhan, China
18007196372@163.com

Abstract. In the field of software security, program structure defects are the main cause of system vulnerability and attack. Especially when the software is very complicated, bug localization and repair become particularly difficult. The aim of the article is to analyze reverse debugging, program structure defect mining, location and repair methods, and enhance the security and stability of software. The binary code is first processed through a combination of several reverse tools, static analysis tools SonarQube and Checkmax to find out potential program structure defects, and then pass the combination of dynamic symbol execution and fuzzy testing method to monitor the behavior of the program during runtime, in order to further verify defects found by static analysis, and to detect new defects generated during runtime. The effectiveness of the method is further evinced by an experiment using 100 open source software. In most cases, the detection rate of static analysis method for structure defects is 78%, while dynamic analysis on the cloud actually confirms that 92% of these defects are true, so reverse debugging fully satisfies the requirement of using the value stored in the register ESP of the local stack to locate the position on the reference source. After passing the run, the pass rate of software in security testing increases from 85% to 95%, indicating that the application of the method can increase the security and reliability of software. The integrated method can locate and repair buggy structure defects of a program, so that the software has fewer security flaws, and better quality.

Keywords: Reverse Analysis · Program Structure · Software Defect Localization and Repair · Dynamic Symbol Execution

1 Introduction

As it is commonly acknowledged, one of the main sources of software vulnerabilities and attacks in modern software security is program structure weaknesses. Given the growing complexity of applications and increasingly changing threat landscape, it illustrates that the more the attackers can exploit the structural weaknesses established for malicious intent. Vulnerabilities resulted in not only the breach of privacy and interrupting the service, but also a damaging blow to user trust and corporate reputation. Therefore, it is especially extra important to identify and remediate the structural weaknesses of the program in a timely and effective manner in order to maintain the security and stability of the software.

The methodology presented in this study combines reverse engineering using a static and dynamic analysis technique, with the aim of helping increase the accuracy and speed of software defect localization. The analysis process begins with a static analysis tool that reviews the binary code of the program in question and identifies any structural defects. In general, static analysis tools can quickly analyze the code and determine if there are any structural vulnerabilities or defects without executing the binary code. After static analysis is completed, dynamic analysis techniques will be employed to observe execution in real time and assess if any structural defects have been identified by the static analysis. Overall, static and dynamic analysis can increase the accuracy of defect localization and expedite repair work through a more comprehensive analysis process.

The first part of the article will clarify the extent to which program structural defects affect software security; the second part will summarize previous research; the third part will explore the specific methods of reverse analysis, static analysis, and dynamic analysis; then, the fourth part will introduce the experimental data and result analysis to confirm the effectiveness of the adopted method; finally, this article will provide research conclusions and propose future work recommendations to provide practical defect localization and repair strategies for software developers and security researchers, in order to improve the overall security and reliability of software.

2 Related Work

Previous researchers have conducted considerable research on software defect localization. Peng Ling proposed a program static defect localization algorithm based on the GL2-DNN model, which combines the global random search ability of genetic algorithms, L2 regularization to prevent model overfitting, and the complex nonlinear learning ability of deep neural networks to address the inconvenience of parameter setting in existing defect localization methods based on deep neural networks [1]. Shen Zongwen retrieved all source codes of real projects through information retrieval to ensure the full utilization of existing features, and then used deep models to mine the semantics of source codes and defect reports to obtain the final localization results. He concluded that through two-stage retrieval, the TosLoc method can quickly locate defects in all code of a single project [2]. Wang Shangwen focused on identifying specific code tokens that cause software defects at a fine-grained level. He established an abstract syntax tree path for code tokens and proposed a fine-grained defect localization model based on pointer neural networks to predict the specific defect code tokens and the specific operational behavior for repairing them [3]. Zhang Zhuo proposed a domain data augmentation method for defect localization models based on adversarial generative networks. This method is based on the model domain (i.e. defect localization spectrum information) rather than the traditional input domain (i.e. program input), and uses adversarial generative networks to synthesize model domain failure test cases that cover the minimum suspicious set, solving the problem of class imbalance from the model domain [4]. Wang Haoren used the sentence coverage information matrix as the original feature set and modeled redundant coverage information as a feature selection problem. He proposed a software defect localization method based on redundant coverage information reduction [5].

Thota M K developed an effective software defect prediction method by using machine learning techniques based on soft computing, which helps predict, optimize features, and effectively learn features [6]. Esteves G proposed a simple model sampling method that uses the minimum feature set to find an accurate model [7]. Wang H proposed a defect prediction method based on gated hierarchical long short-term memory networks, which uses a hierarchical LSTM network to extract semantic features from word embeddings in abstract syntax trees of source code files, and utilizes traditional features provided by the PROMISE knowledge base [8]. Goyal S aimed to reduce software development costs by focusing testing efforts on the predicted faulty modules [9]. Ardimento P proposed a new method based on a large feature set containing product and process software metrics extracted from the submission and evolution of software projects [10]. These studies have provided some assistance for the work of this article, and this article will continue to explore software defect localization in more depth through reverse analysis techniques.

3 Method

3.1 Application of Reverse Analysis Techniques

In this article, reverse analysis will be used to repair software, which includes two processes: disassembly and decompilation. Disassembly is the process of converting binary code into assembly language code for the purpose of analyzing the instruction flow of a program. Decompilation seeks to convert assembly language code back into high-level language code and has importance, it can help to ascertain the overall logic of the program. The first step in performing reverse analysis is to utilize the disassembly tool IDA Pro to convert the binary code of the target executable program to assembly code, which will produce a set of assembly instructions to logically express control flow relations between sections in the program code along with data flow relationships that involve execution states. With these assembly codes, the basic execution blocks can be determined, and their relationships can be expressed through a control flow diagram to convey an understanding of the program execution path. In turn, after analyzing the control flow, data flow can be analyzed to identify the usage of variables and registers to locate possible defect locations. If we consider the output of a function with x as input and $f(x)$ as output, then f needs to be further analyzed for its implementation details to reveal possible defects or weaknesses [11]. In addition, runtime information can be supplemented through a combination of dynamic analysis and static analysis, which involves collecting information during program runtime and then associating it with the analysis results. This process can reveal execution specific defects, such as memory leaks or uninitialized variables. During this process, pattern matching techniques may also be involved in identifying known defect patterns or malicious code fragments in wireless communication. Over time, researchers may establish a defect feature library, which can then be used for more automated tool based defect scanning of appropriate source code and form a fast feedback loop [12].

3.2 Use of Static Analysis Tools

The process of static analysis involves analyzing source code or binary files to identify potential defects and security vulnerabilities without executing the program. Static analysis tools can automate code inspection through various techniques, thereby improving security and code quality. Through static code inspection, common defects in the code are identified [13], such as uninitialized variables, null pointer references, and array out of bounds, which directly affect the stability and security of the program. Tools typically use Abstract Syntax Tree (AST) or Control Flow Graph (CFG) to analyze program structure. For example, given a variable in the code, static analysis tools can track its usage and identify potential defects by constructing a control flow diagram. Secondly, static analysis tools can also detect security vulnerabilities, such as SQL injection. The tool identifies unsafe code patterns through pattern matching and data flow analysis, and can check whether user input has been properly validated and escaped, thereby determining whether there are potential injection vulnerabilities.

This article uses SonarQube and Checkmax to scan the binary code of the program and identify potential structural defects. When conducting static analysis, tools typically use some metrics to evaluate the quality and security of the code. The common metric is code complexity, which is calculated using the following formula:

$$CC = \frac{E - N + 2P}{P} \quad (1)$$

CC is the complexity of the circle, E is the number of edges in the program, N is the number of nodes, and P is the connected component in the program. A higher loop complexity usually means a more complex code structure, and potential defects and vulnerabilities are more difficult to identify. Static analysis tools also provide reporting functionality, generating detailed lists of defects and vulnerabilities to help understand the nature and severity of problems. The report usually includes the type, location, and recommended repair methods of the defect, which can help to quickly locate the problem and make repairs.

3.3 Integration of Dynamic Analysis Techniques

This article uses dynamic analysis techniques to integrate different types of dynamic analysis methods to improve the effectiveness of system analysis and performance optimization. The main goal of this comprehensive approach is to overcome the limitations of a single dynamic analysis technique, in order to gain a more comprehensive understanding and improve the behavior of software systems. It includes but is not limited to dynamic symbol execution, fuzz testing, dynamic slicing, and dynamic memory analysis [14]. Dynamic symbol execution explores all possible execution paths of the program through symbol execution paths to detect potential vulnerabilities and errors. The following paths can be explored:

$$Path = \{p1, p2, p3, p4, ..., pn\} \quad (2)$$

P represents the execution path, and fuzzy testing discovers boundary conditions and abnormal behavior in the program by inputting random or semi random data. Dynamic

slicing technology can help developers understand the scope of influence of specific variables or statements in a program. Dynamic memory analysis is used to detect memory leaks and potential memory errors. Integrating these dynamic analysis techniques together can generate synergies and enhance a comprehensive understanding of software system behavior. Combining dynamic symbol execution and fuzz testing can more effectively discover vulnerabilities in programs, and the combination of dynamic slicing and dynamic memory analysis can help developers quickly locate and fix defects in programs [15]. Table 1 shows a list of identified defects:

Table 1. Identified defects

Test ID	Input Data	Execution Path	Detected Defect Type	Description
T001	Input Set A	Path 1	Memory Leak	Detected unreleased memory
T002	Input Set B	Path 2	Resource Race	Conflict in thread access
T003	Random Input C	Path 3	Buffer Overflow	Input exceeds buffer size

4 Results and Discussion

4.1 Test Environment

In software security reverse analysis, it is necessary to configure a high-performance computer with 16GB of memory and a quad core processor to support the operation of complex static and dynamic analysis tools. Using SSD hard drives to improve data read and write speeds, ensuring no bottlenecks during large-scale code analysis, while installing the Windows operating system to ensure compatibility with different types of applications. The development environment should be equipped with an integrated development environment Eclipse for editing and debugging source code, configuring dynamic analysis tools Valgrind and fuzz testing tools AFL, as well as static analysis tools SonarQube and Checkmark, to achieve comprehensive security analysis. In the testing platform, configuring virtual machines or Docker containers to isolate the analysis environment, prevent any impact on the main system, and ensure the security and controllability of the experimental process. Finally, building a version control system to manage code versions, facilitating the tracking of defect repair processes and retrospective analysis, thereby improving the efficiency and accuracy of reverse analysis.

4.2 Dataset Display and Testing Methods

Bugzilla is an open-source defect tracking system used to manage defects and issues in software projects. It was originally developed by Mozilla and has now been widely

used in multiple open source and commercial projects. Bugzilla provides a range of features to help teams effectively record, track, and manage defects. Table 2 shows a partial display of the dataset:

Table 2. Bugzilla partial display

ID	Title	Status	Severity	Priority	Assigned To	Creation Date	Update Date
123	Page loading slowly	Confirmed	Medium	High	Alice	2024-01-15	2024-01-20
124	Application crashes on export	New	Severe	Highest	Bob	2024-01-16	2024-01-16
125	Image display incorrect	Resolved	Minor	Medium	Charlie	2024-01-17	2024-01-18
126	Missing user manual	New	Medium	Low	None	2024-01-18	2024-01-19
127	Login page fails to load	Closed	Severe	High	Alice	2024-01-19	2024-01-21

Selecting 100 open-source software projects from the dataset for testing, and use the Sober algorithm to compare the two probability model methods with the reverse analysis method proposed in this paper. Firstly, identifying structural defects, then further validate the identified defects, and finally perform software repair.

4.3 Test Results

Figure 1 shows the software defect identification results:

After testing 100 open source software projects, the reverse analysis method in this article performs well in identifying structural defects, achieving a recognition rate of 78%, proving the deep understanding of software structure and sensitivity to abnormal patterns by the reverse analysis method. The recognition rate of the Sober algorithm is 62%, and the probability model is 72%, indicating a lower recognition efficiency. The Sober algorithm faces limitations in its applicability, resulting in poor performance in structural defect recognition, while probabilistic models are affected by model complexity and training data quality. The high recognition rate of the reverse analysis method in this article provides important support for software quality assurance, which can help development teams discover and solve potential problems in a timely manner, and improve software reliability.

Continuing to verify the authenticity of the identified defects, dynamic analysis is used to validate the identified defects. The validation results are shown in Fig. 2:

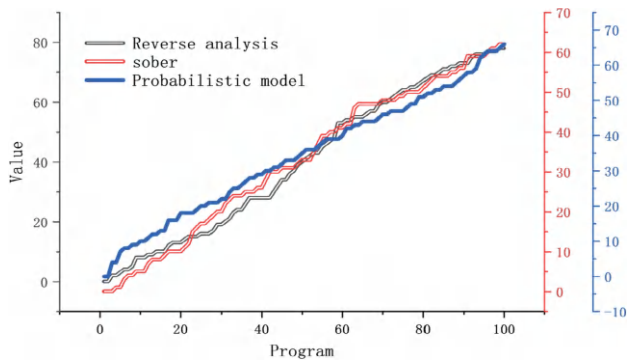


Fig. 1. Defect identification quantity

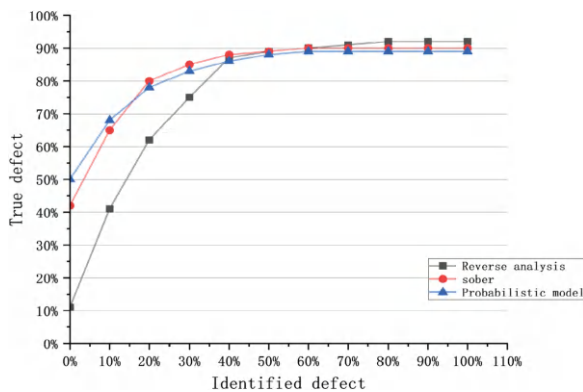


Fig. 2. Verification results

Observing the above results, it can be observed that reverse analysis, Sober, and probability models exhibit different trends and accuracies in identifying detected defects. Among them, the defect recognition rate of reverse analysis gradually increases with the severity of defects, but the growth slows down when the defect severity is above 0.6. It verifies 92% of the identified defects as real defects. The Sober method has a high recognition rate at lower defect severity levels, but its recognition ability slightly decreases as the defect severity increases. The probability model exhibits relatively stable performance across the entire range and can maintain a high recognition rate for defects of different severity levels. This article conducts reverse analysis to detect defects by deeply analyzing the execution path and data flow of the code, thus enabling more accurate problem localization in more severe defect situations. Finally, three methods are used to repair the program, and the software qualification rates before and after repair are compared. The comparison results are shown in Fig. 3:

The () in the figure represents the pass rate before repair. By using three methods of reverse analysis, Sober, and probability model to repair the program, it can be clearly seen that the pass rate of the software has changed before and after repair. In terms of

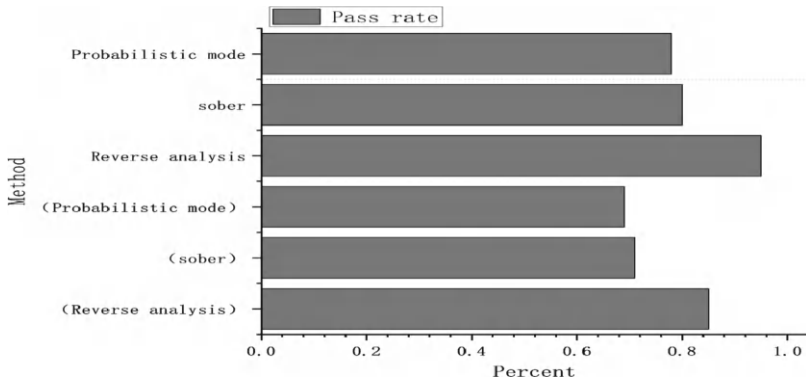


Fig. 3. Program qualification rate

reverse analysis, the software qualification rate before repair is 0.85, but after repair, the qualification rate increases to 0.95, showing significant improvement. The Sober method has a software qualification rate of 0.71 before repair, which increases to 0.8 after repair and shows a certain improvement. The probability model has a software qualification rate of 0.69 before repair, which slightly increases to 0.78 after repair. Although the improvement is small, it still shows a trend of improvement.

5 Conclusion

The key to locating program structural defects lies in a deep understanding of the program's execution path and data flow, and reverse analysis techniques can effectively help locate these defects. The repair method requires targeted modification and optimization of the program based on the results of reverse analysis, in order to eliminate potential security risks and vulnerabilities. Through testing, it was found that the pass rate of the software repaired by reverse analysis in security testing has increased from 85% to 95%. The application of this method significantly improves the security and reliability of the software. The research on software security reverse analysis in locating and repairing program structural defects provides important ideas and methods for improving software security. Future development will pay more attention to technological innovation and cross-border cooperation to address increasingly complex security challenges.

References

1. Peng, L., Liu, Z., Peng, M.: A program defect localization method based on GL_ (2) - DNN for statement coverage. *Comp. Appl. Soft.* **40**(1), 46–52+155 (2023)
2. Shen, Z., et al.: Software defect localization method integrating information retrieval and deep model features. *Journal of Software* **35**(7), 3245–3264 (2024)
3. Wang, S., et al.: Fine grained defect localization based on pointer neural network. *Journal of Software* **35**(4), 1841–1860(2024)

4. Zhuo, Z., Yan, L., Xiaoguang, M., Jianxin, X., Xi, C.: Domain data augmentation method for defect localization model based on adversarial generative network. *Journal of Software* **35**(5), 2289–2306 (2024)
5. Haoren, W., Zhanqi, C., Lei, Y., Xiang, C., Liwei, Z.: A software defect localization method based on redundant coverage information reduction. *Acta Sinica* **52**(1), 324–337 (2024)
6. Thota, M.K., Shajin, F.H., Rajesh, P.: Survey on software defect prediction techniques. *Int. J. Appl. Sci. Eng.* **17**(4), 331–344 (2020)
7. Esteves, G., Figueiredo, E., Veloso, A., et al.: Understanding machine learning software defect predictions. *Autom. Softw. Eng.* **27**(3), 369–392 (2020)
8. Wang, H., Zhuang, W., Zhang, X.: Software defect prediction based on gated hierarchical LSTMs. *IEEE Trans. Reliab.* **70**(2), 711–727 (2021)
9. Goyal, S.: Handling class-imbalance with KNN (neighbourhood) under-sampling for software defect prediction. *Artif. Intell. Rev.* **55**(3), 2023–2064 (2022)
10. Ardimento, P., Aversano, L., Bernardi, M.L., et al.: Just-in-time software defect prediction using deep temporal convolutional networks. *Neural Comput. Appl.* **34**(5), 3981–4001 (2022)
11. Xu, J., Ai, J., Liu, J., et al.: ACGDP: An augmented code graph-based system for software defect prediction. *IEEE Trans. Reliab.* **71**(2), 850–864 (2022)
12. Chen, L., Wang, C., Song, S.: Software defect prediction based on nested-stacking and heterogeneous feature selection. *Complex & Intelligent Systems* **8**(4), 3333–3348 (2022)
13. Xiaolong, X.U., Wen, C., Xinheng, W.: RFC: a feature selection algorithm for software defect prediction. *J. Syst. Eng. Electron.* **32**(2), 389–398 (2021)
14. Verna, E., Genta, G., Galetto, M., et al.: Inspection planning by defect prediction models and inspection strategy maps. *Prod. Eng. Res. Devel.* **15**(6), 897–915 (2021)
15. Malhotra, R., Jain, J.: Predicting Software Defects for Object-Oriented Software Using Search-based Techniques. *Int. J. Software Eng. Knowl. Eng.* **31**(02), 193–215 (2021)



Application of Data Mining in the Development and Management of Software Engineering in Cloud Computing Platform

Qing Tan(✉)

School of Information Technology, Luoyang Normal University, Luoyang, China
edutanqing@163.com

Abstract. Data Mining is a process of extracting valuable information and knowledge that people have not observed from a large number of complicated data. Massive data mining based on cloud computing can optimize the dynamic resource scheduling and allocation of information. The application of data mining technology in computer software engineering can quickly find out the data which is beneficial to software engineering management, thus providing useful information for software development and management. Data mining technology has significant application advantages in software vulnerability detection, fault repair, source code development and other fields. This paper proposes the application of data mining technology to software engineering on the basis of cloud computer. The experimental results show that the data mining technology proposed in this paper can effectively improve the quality of software engineering development and management under the cloud computing platform.

Keywords: Data mining · Cloud computing · Software engineering · Data preprocessing · Software vulnerability

1 Introduction

The current network data has been growing exponentially, in front of a large amount of data, how to find efficient data processing technology, methods and means to dig out effective information, has become an urgent need for society and enterprises. However, there are some defects and deficiencies in the existing data mining methods in the practical application, not only the information gain value of the mining results is relatively low, but also the data mining is time-consuming, which cannot meet the actual needs.

Most of the hidden information can be used, valuable and potential. From the commercial point of view, data mining refers to the process of analyzing and processing business activity information, the purpose is to realize the integration, transformation, extraction and analysis of business data, to obtain business information that is conducive to financial management and the development of the company's operation, and to provide financial data support for management decision-making.

Distributed data mining technology is to build a distributed computing environment based on the Internet. Its computing process mainly relies on dynamic and scalable

virtualization resources to mine distributed data in the network space. Users cannot directly control the computer, and dynamic virtualization resources combine global information by integrating local information.

Data mining technology is a method of discovering, extracting and analyzing potential information and hidden patterns in large-scale data, which integrates the knowledge and technology of machine learning, statistics, database technology and artificial intelligence.

Data mining technology provides a powerful tool for software engineering, which can play a key role in requirements analysis and user behavior prediction. By analyzing user historical data, behavior patterns, and feedback, development teams can better understand user needs, predict user behavior, and optimize software functionality and user experience. This data-driven requirement analysis can help developers avoid unnecessary feature development, save time and resources, and ensure that the developed software is closer to the needs of users. Each algorithm plays a different role in archives management.

In this paper, the membership inference attack based on data mining technology on genomic data can accurately determine whether a specific object individual belongs to a specific disease [1]. The author [2] uses data mining technology to give full play to the user's motion characteristics and rules, and can restore the target's motion track from the aggregated motion track without any prior knowledge, so as to achieve the purpose of reasoning about the target. The authors [3] capture characteristic keywords on the basis of collecting user behavior data, and then use data mining to find behavioral data such as total online time and total traffic in online communities. G Izmirlian [4] used Random Forest (RF) algorithm to conduct related research on SELDI-TOF proteomics, and especially emphasized its application in cancer prevention.

In this paper, the characteristics of information user behavior are mined according to the time characteristics, and the starting sequence vector is obtained by calculating the average data in the sliding window, and then the user behavior is divided into several equal time slices. The user behavior is counted by sampling, and the average query frequency is used as the index to extract the characteristics of user query behavior. The feature calculation process is too single, which is easy to lead to large errors [5]. In this paper, the decision tree C4.5 algorithm was proposed to classify the lithology of remote sensing images in the Three Gorges reservoir area [6], and the classification results of C4.5 algorithm were compared with K-means clustering algorithm in this study. This paper presents a new decision tree pruning algorithm, which has some practical value [7]. In this paper, a wrapper feature selection algorithm based on random forest (RFFS) is proposed, which uses random forest algorithm as the basic tool [8].

The data pre-processing phase includes data cleaning, integration, transformation and specification to ensure the quality and consistency of data. Pattern discovery is the core step, which uses algorithms such as classification, clustering, and association rules to discover patterns and regularities in data. This paper proposes the application of data mining technology to software engineering on the basis of cloud computer.

2 Related Work

The application of these advanced technologies can reduce the workload of financial managers to a certain extent, so as to improve the efficiency of financial management.

Parallel data mining technology is a data mining working mode based on parallel algorithms [9]. Its working principle is to coordinate processes through different algorithm process sets that can be executed at the same time, so that many processors can filter out creative and valuable information.

The technical application flexibility of cloud computing platform is relatively high, and the support of software and hardware virtualization technology makes the system multi-accommodative of cloud computing platform steadily improved. Because the cloud computing platform can carry out efficient data transmission based on the compatibility of multiple systems, some low-performance system devices can also carry out multi-functional system expansion, so as to realize the dynamic management of iot data mining information in a single system.

(1) Data preprocessing. Before data mining, the original data should be cleaned and transformed, including data cleaning, data integration, data transformation and data reduction, in order to eliminate noise, missing values and redundancy in the data; (2) Feature selection and extraction. In the process of data mining, selecting and extracting appropriate features is a very critical step. In this step, through feature selection and feature extraction technology, the most relevant or representative features are selected from the original data, so as to facilitate the subsequent data modeling and analysis.

Model building is the core of data mining. In this order segment, the appropriate data mining algorithm is selected according to the goal of the task, such as classification, clustering, association rule mining, etc. Through the selection algorithm, the preprocessed data is fed into the model to train a model that ADAPTS to the characteristics of the data [10]. During model training, model parameters may need to be tuned to achieve optimal performance. After the training is completed, the model can be used to perform prediction, classification, clustering and other operations on new data to obtain useful information [11]. The results of the model construction can help in decision-making, predicting future trends, discovering regularities, etc.

$$u_n = K_P \left[e_n + \frac{1}{T_I} \sum_{i=0}^n T + T_D \frac{e_n - e_{n-1}}{T} \right] \quad (1)$$

In Eq. (1), u represents the interval applicability for data feature g ; K shows the overall distribution level of data feature g . e denotes the data distribution level of a certain partition of the data feature g . The applicability I_g can reflect the degree of correlation between two sets of data.

Data mining is extracting useful information and patterns from preprocessed data. In telecommunication industry, the methods of data mining mainly include classification, clustering, association rules, sequential patterns and anomaly detection. Among them, classification and clustering are the two most commonly used methods, classification is to classify the data according to the known classification criteria, clustering is to group the data naturally, association rules are to discover the relationship between the

data, sequential patterns are to discover the temporal sequential patterns of the data, and anomaly detection is to find the abnormal patterns of the data.

Through the application of data mining technology in marketing analysis, enterprises can better understand the market and consumers, accurately grasp the market opportunities, and provide personalized products and services, thereby promoting market growth and competitive advantage. These three processes are coordinated and progressive, as is shown by Fig. 1, which can effectively classify data information.

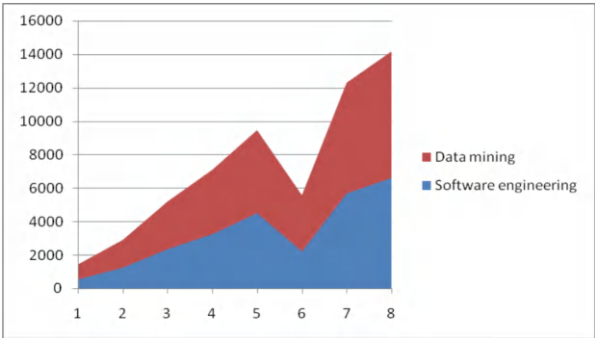


Fig. 1. The comparison result of data mining technology and software engineering in the cloud computing platform

Strengthening data mining technology in computer software engineering has an important significance and demand; the staff needs to combine software development goals and requirements to carry out data mining work. Because to complete the development of the software, it is necessary to edit the source code, so the staff needs to combine the goals and requirements of the software to do a good job in the development of the source code.

As a new and high technology, data mining integrates the technological advantages of computer technology, big data technology and machine learning, and has many theoretical supports [12]. The application of data mining technology in precision marketing strategy can carry out customer identification, strategy formulation and other activities in the form of mathematical models, improve the reliability of quantitative analysis of precision marketing, and make marketing strategy more scientific [13]. In addition, data mining technology can deeply analyze the past precision marketing strategy, and provide scientific suggestions for the formulation of the next precision marketing strategy for enterprises by analyzing the factors of success and failure of marketing strategy, and enhance the reliability of the strategy.

Data mining aims to discover patterns in data, which can be regular, abnormal, repetitive, trend, etc. By identifying these patterns, organizations can extract valuable information from the data. Data mining can be used to build predictive models that enable organizations to predict future trends and outcomes based on past data, which is essential for business decision making and planning. Classification is the grouping of data into different categories or labels, while clustering is the grouping of data into similar groups, both techniques help organizations better understand their data and identify patterns.

Online social media data mining is a complex and challenging task, which uses artificial intelligence and database technology to obtain, analyze and mine valuable information from massive social media platforms. In this process, data is regarded as the precious soil, and cloud platform is regarded as the infrastructure that hosts the data and mining algorithms.

3 Application of Data Mining to Software Engineering on the Basis of Cloud Computer

The full name of network data mining service is neural network system model, the main technical principle is to use cloud computing technology to establish a neural network mathematical model, based on the processing unit similar to human brain neurons can be interconnected nodes, data mining process based on the input data of these nodes for service, processing and other information decisions.

By optimizing the statistical computing ability of distributed computer equipment, cloud computing platform realizes the integration of multivariate data processing resources, and realizes the sharing of data processing resources relying on virtual cloud service platform, so that physical data resources can rely on cloud computing technology to achieve a balanced allocation [14]. Compared with the traditional data computing mode based on a single server system, the cloud computing platform reduces the threshold of data information calculation through the pooling of multiple different server equipment resources, as is shown by Table 1, so that individuals and enterprise users do not need to build a server management platform to effectively utilize the data resources of the Internet of Things.

Table 1. Application of data mining to software engineering on the basis of cloud computer

Data mining algorithms	Ac	Pre	Rec	F1	AUC
Apriori	0.25	0.86	0.54	1.28	0.87
K-means	0.28	0.82	0.62	1.36	0.86
C4.5	0.36	0.74	0.71	1.24	0.82
CART	0.14	0.69	0.68	1.33	0.74
SVM	0.42	0.64	0.62	1.56	0.69
KNN	0.29	0.92	0.63	0.93	0.95
EM	0.56	0.59	0.58	2.10	0.63
ID3	0.11	0.81	0.81	1.96	0.85

After completing the application of the data mining algorithm, the model needs to be evaluated and optimized. The model can be evaluated by comparing the prediction accuracy, callback rate, precision rate and other indicators, and according to the evaluation results, the algorithm can be optimized. These steps form the core of the data mining

process, valuable information and knowledge are finally obtained to provide support and guidance for decision making.

Firstly, the mining of software version information can help the development team to analyze and predict the trend of software evolution. The patterns and rules of software evolution can be found by mining data such as change logs of historical versions, code changes and comments from developers. This helps the team predict which modules may need to be changed more frequently, which modules may face performance issues, and which features may need to be extended or optimized.

Secondly, software version information mining can help identify and manage code defects. By analyzing the bug reports, code changes, test records and other data in historical versions, a defect prediction model can be built to predict the possible defects in future versions.

Data mining algorithm is the core of data mining, which mainly includes four common mining methods: classification, clustering, prediction and association rules. Classification involves grouping data into different categories. Clustering is the grouping of similar data objects. Forecasting is to predict the future trend based on historical data. Association rules are about finding association patterns in the data.

Data mining techniques can provide personalized product recommendations for customers through the analysis of customer behavior and preferences. By mining customers' purchase history, browsing behavior, interests and other data, personalized recommendation system can identify potential cross-selling opportunities and personalized purchase preferences, provide customers with accurate and personalized product recommendation, as is shown by Fig. 2, and improve customers' purchase satisfaction and loyalty. Data mining technology can predict the future needs and behaviors of customers by analyzing and modeling customer data.

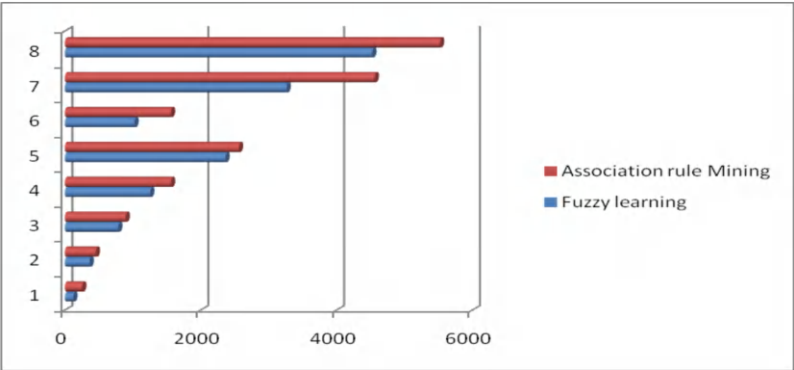


Fig. 2. The comparison result of association rule mining and fuzzy learning in the cloud computing platform

In Eq. (2), and LBP represent two unknown parameters of the line respectively; s is the number of lines. In order to find out the best classification mining line, two unknown parameters need to be determined, and a part of the data set is used as training samples to establish a training set. The training set is input into the support vector machine (SVM)

through training and learning, and the linear regression of the data in the training set is realized.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \quad (2)$$

In order to accurately analyze user preferences, operating habits and other information, technical personnel need to apply the information input, search and analysis functions in data mining technology, process user software application information records, analyze and summarize the rules, and improve the quality of software development. It should be noted that during the application of data mining technology, technicians need to clarify the object of data mining, and classify, summarize and evaluate the final data mining results. In the early stage, the data mining method should be adjusted and optimized according to the data characteristics and the basic requirements of software project management.

With the help of data mining technology, the clone of program code can be realized, thus reducing the workload of staff and improving work efficiency. In the development of computer software engineering, the number of program codes that need to be developed is massive, and staff need to spend a lot of time and energy to do a good job in code entry, which is easy to affect the overall development efficiency.

The essence of data mining needs the support of massive data and key technologies to explore effective information that fits the needs of enterprise development. Traditional marketing thinking has the characteristics of heavy profit, high cost and many conditions, which is not consistent with the inner core of data mining technology. First of all, the core focus of traditional marketing strategy is the expansion of corporate profits. In order to improve revenue, the analysis of customer data is obviously insufficient, and the essential reason for profit improvement is not explored, so it cannot achieve sustainable development, which is contrary to the emphasis on relevance of data mining technology.

Data mining can be used to analyze supplier performance data, including delivery punctuality, product quality, cost, and service level. By monitoring these key indicators, enterprises can identify high-performing suppliers and potential problem suppliers, and make corresponding decisions, such as supplier performance reward or supplier replacement.

Cloud computing technology is a specific application of big data technology, which tends to develop in conjunction with cloud storage technology, big data management technology and other related technologies. Whenever one of the technologies shows a trend of development, other related technologies also tend to develop forward in a matching manner, and finally realize the overall progress of big data technology.

4 Results and Discussion

The data collection of Internet of things data mining on cloud computing platform needs to collect data information from various data sources and store it in a central location or database. Among them, due to the differences in data density and data integrity of different types of data information content, data acquisition needs to be optimized by

combining data quality control and data correction. Sensor calibration and data noise processing are used to improve the accuracy and effectiveness of data acquisition. To ensure that the data collected at each stage has a certain application value.

In the experiment, the information gain value was used as the performance comparison evaluation index of the three methods, and the information gain value could reflect the value and effect of the characteristic information of data mining on user churn prediction. The larger the information gain value, the greater the value of the mining results for user churn prediction. The experiment takes the amount of data mined as the variable and 3000Byte data as the base. After completing data mining, 2000Byte is added to the original data until the amount of data mined reaches 50000Byte.

The association function of data mining technology can summarize the information that appears in the database at the same time, and find the regular relationship of these data information. For example, by mining and analyzing sales data, we can find out the factors that will affect the sales effect, as is shown by Eq. (3) and develop more reasonable sales strategies and programs.

$$x^{(1)}(t) = \left(x_1^{(0)} - \frac{u}{a}\right)e^{-a(t-1)} + \frac{u}{a} \quad (3)$$

The platform provided by cloud computing technology is an ideal environment for data mining, the greater the amount of data, the more accurate the depth and results of the mining, its high precision data mining results make it have a strong information value, which is conducive to optimizing the Internet information environment, and discovering the true situation hidden behind the data. Therefore, massive data mining based on cloud computing must have a sufficient amount of information resources, the greater the amount of information, the more conducive to cloud computing technology to adopt scientific algorithms, and the better to ensure the accuracy of data analysis results.

The Internet of things data mining on cloud computing platform mainly uses two modes of real-time storage and batch storage to write data. The advantage of batch storage is that it can use a relatively short period of time to complete a large number of data information storage tasks, but it has certain requirements for the system's sequential writing performance. Relying on the server device of enterprise-class NAS storage system, it can provide efficient data read and write support for IOT devices to generate a large amount of data, as is shown by Table 2.

It can be seen from the data in the Table 2 that the information gain value mined by the design method in this experiment is relatively higher, indicating that the information mined by the design method is more valuable for user churn prediction. The data mining amount is also taken as the variable, and the time consumption of the three methods under different data mining amounts is counted, and the experimental data is recorded by electronic table.

Mining software execution records is also very helpful for fault detection and troubleshooting. By monitoring logs, error messages, and the exception stack, we can trace and locate potential problem sources, so that we can find and resolve software failures in time. This helps to improve the stability and reliability of the software, and reduce user complaints and losses caused by faults.

Technicians analyze the actual situation of each slice fault through data mining technology, and construct data traces as a reference basis for fault detection and repair.

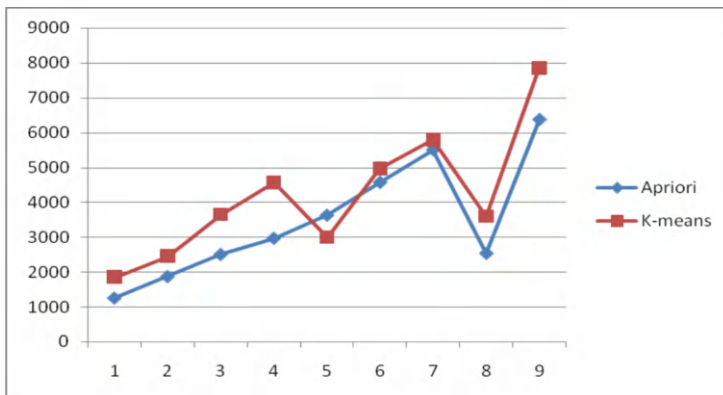
Table 2. Experimental data analysis of data mining for software engineering management under cloud platform

Data mining algorithms	Apriori	K-means	C4.5	CART
RMSE	0.456	0.896	0.711	0.802
Purity	0.658	0.756	0.523	0.711
Mutual Information	0.854	0.963	0.452	0.806
F1	0.741	0.841	0.633	0.636
MSE	0.363	0.633	0.412	0.523
Accuracy	0.585	0.756	0.622	0.712
Precision	0.639	0.866	0.741	0.816
Recall	0.562	0.856	0.632	0.703

For example, after the technicians detect the data program slice, they can locate the slice according to the data mining rules, so as to determine the scope of fault detection and ensure the effectiveness of fault detection and repair in the software.

Staff can use data mining technology to do a good job of extracting and mastering code components. In computer software engineering, the main function of the code component is to retrieve the structure of the program code. At present, the program code of the computer software is generally in the format of characters.

Cloud computing technology is the core content of big data technology, data mining work in addition to accuracy, convenience, but also adhering to the principle of security to carry out. It is a guiding principle that must be followed when performing data mining, as is shown by Fig. 3.

**Fig. 3.** The comparison result of Apriori and K-means in the cloud computing platform

The design of an integrated system for data analysis should integrate the analysis results into different application devices to support decision-making and automation

processes, and fully provide value-added services for personalized data services. Generally, the design of integrated system needs to use API program interface, embed the program into the system function, and then use third-party applications similar to ERP system and CRM system for data relationship.

In the process of software vulnerability testing, technicians use data mining technology to record the test content in detail, and select the most effective test method from it. In addition, technicians need to apply data mining technology to deal with vulnerability data, especially the analysis and extraction of redundant data, and screen out valuable data as backup from massive information. The valuable data obtained by using data mining technology in computer software vulnerability scanning can be used as the basis for creating data models.

Cloud computing technology collects a large number of users' personal real information when mining Internet information. The disclosure of these user information is easy to lead to the loss of people's personal property, causing security cases and widespread social concern. Therefore, we must pay great attention to the system security of cloud computing platform while applying cloud computing technology for data mining.

The data storage logic of IoT data mining on cloud computing platform is basically the same as that of ordinary computer data storage, but the technical characteristics are completely different. In terms of data storage types, mainstream databases include NoSQL databases, distributed file systems, and data lakes. The nature and storage cycle of data determine the application requirements for different types of database systems.

5 Conclusion

Cloud computing technology is not a new technology strictly speaking, but due to the explosion trend of Internet information data, in order to better serve the needs of social, based on cloud storage technology, using the cloud computing platform's own characteristics of dynamic resource scheduling, high virtualization and high availability to improve the efficiency of log data analysis and processing, and reduce the cost of data mining work. The traditional data mining methods are optimized and improved, which well meets the growing demand of social economy for information services and maintains a good development trend.

Internet of things data mining based on cloud computing platform can improve the efficiency of data information processing and the integration ability of all kinds of data information content in the new era, further strengthen the ability of data mining technical services to provide data application support for enterprise-level customers, and effectively meet the application requirements for all kinds of different data information content in different scenarios. It lays a solid foundation for data mining and data application management system.

Data mining technology is a new technical means launched in the era of big data, which has the advantages of deep mining data and efficient data analysis. According to the fundamental needs of computer software engineering, technical personnel need to prepare the application plan and scheme of data mining technology, make full use of the advantages of this technology, extract the real application value of data, and provide support for the construction of computer software engineering.

In short, in the process of computer software engineering development, strengthening the use of data mining technology can effectively do a good job in the management of software information, do a good job in the analysis of program code, do a good job in the detection of software faults, and so on, and play an important role in promoting the overall quality of software development.

References

1. Kao, S.C., Fulham, M., Wong, K., et al.: Application of data mining technology in computer network virus defense. *Information and Computer* **35**(10), 43–45 (2023)
2. Ruby, A.U., Theerthagiri, P., Jacob, I.J., et al.: Function Extraction Based on CFPS and Digital Financial Index: Data Mining Techniques for Prognosis of Operational Risks of Financial Institutions. *Journal of Sensors* (2022)
3. Feyyad, U.M.: Data mining and knowledge discovery: making sense out of data. *IEEE expert* **11**(5), 20–25 (2020)
4. Izmirlian, G.: Discussion on effective combination and application of sensor technology based on big data, cloud computing and Internet of Things. *Network Sec. Technol. Appl.* **10**, 168–170 (2023)
5. Ernst, N.A., Baldassarre, M.T.: Registered reports in software engineering. *Empir Software Eng.* **28**, 55 (2023)
6. Bhattacharya, A., Baweja, T., Karri, S.P.K.: Application of data mining technology in computer software engineering. *Digi. Technol. Appl.* **41**(1), 126–128 (2023)
7. Cassee, N., Agaronian, A., Constantinou, E., et al.: Transformers and meta-tokenization in sentiment analysis for software engineering. *Empir Software Eng.* **29**, 77 (2024)
8. Aparicio-Morales, Á.M., Moguel, E., Bibbo, L.M., et al.: An overview of quantum software engineering in Latin America. *Quantum Inf. Process.* **23**, 380 (2024)
9. Guo, C.: Computer Data Mining Technology and application. *China Manage. Info.* **25**(4), 178–180 (2012)
10. Sarkar, A.: Automated quantum software engineering. *Autom. Softw. Eng.* **31**, 36 (2024)
11. Song, Y.: Research on application of data mining technology in software development. *J. Softw.* **42**(9), 158–160 (2021)
12. Hongsheng, X., et al.: Dynamic SFC placement scheme with parallelized SFCs and reuse of initialized VNFs: An A3C-based DRL approach. *J. King Saud Univ. Comp. Info. Sci.* **35**(6), 101577 (2023)
13. Jiwu, Y.: Data mining technology from the perspective of cloud computing. *Electr. Technol. Softw. Eng.* **5**, 151 (2019)
14. Dwivedi, K., Haghparast, M., Mikkonen, T.: Quantum software engineering and quantum software development lifecycle: a survey. *Cluster Comput* **27**, 7127–7145 (2024)



Credit Rating Optimization Model Based on Deep Q-Network

Yijiao Fan(✉)

JP Morgan, New York 11101, United States
yijiaofanedit@outlook.com

Abstract. With the advancement of information technology, the application of blockchain technology in various fields continues to increase. However, the information isolation between different blockchain networks leads to the phenomenon of information silos, which impedes the exchange of information and the flow of value between networks. In order to solve this problem, cross-chain technology comes into being, among which notary mechanism is widely concerned because of its easy implementation and good scalability. However, the current notary public mechanism is faced with the problem of insufficient reliability of notary public, which limits the popularization of its practical application. This paper proposes a credit score optimization method based on weighted leader ranking algorithm to improve the performance of cross-chain notary mechanism. Firstly, a new trust evaluation model is constructed, which can comprehensively evaluate the behavior of notary public by considering multiple evaluation indexes according to the characteristics of cross-chain transaction of notary public mechanism. Experimental results demonstrate that this model significantly outperforms traditional methods in evaluation performance. It is then applied to the notary mechanism, leading to the proposal of a cross-chain notary mechanism that effectively quantifies notary node behavior, thereby enhancing reliability. The model shows strong performance in selecting trusted notary nodes. Finally, to address the centralization issue in notary elections, a deep Q network-based election algorithm is introduced, ensuring both cross-chain transaction performance and election fairness, while mitigating centralization tendencies.

Keywords: Cross-Chain Technology · Notary Mechanism · Trust Evaluation Model · Weighted Leader Ranking Algorithm · Deep Q Network (DQN)

1 Introduction

In October 2008, Satoshi Nakamoto published Bitcoin: A Peer-to-Peer Electronic Cash System, marking the birth of blockchain technology and leading the decentralized financial revolution. As a distributed ledger, blockchain technology jointly maintains transaction data through network nodes, and uses cryptography technology to ensure that the data is immutable and unfalsifiable, demonstrating its core advantages of decentralization, immutable and traceable. These characteristics give blockchain significant potential in terms of data security and transparency. However, despite the initial success

of blockchain technology in the financial sector, its development in other application scenarios still faces many challenges.

The evolution of blockchain technology has gone through the process from version 1.0 of the digital currency stage, to version 2.0 of the smart contract application, and then to version 3.0 of the wide application in industry, the Internet of things, logistics and other fields. Blockchain 1.0 mainly focuses on digital currencies such as bitcoin, while 2.0 introduces smart contracts and expands the application of the technology to the financial field. Phase 3.0 explores the application of blockchain technology in a wider range of fields, promoting the transformation of the Internet of Information into the Internet of value. However, the cross-chain capabilities of this technology still have significant limitations, that is, the flow of data and value between different blockchain networks, creating information silos.

In order to deal with these challenges, this paper proposes a notary mechanism optimization scheme based on weighted leader ranking algorithm. By constructing a new trust evaluation model, the behavior of notary public can be evaluated more comprehensively and its reliability can be improved. At the same time, the deep Q network (DQN) algorithm is used to improve the notary election process, so as to reduce the problem of election centralization and improve the cross-chain transaction performance of the notary mechanism. These innovative studies not only provide new solutions for the optimization of the notary mechanism, but also help to promote the practical application of blockchain technology and industrial integration, and promote the overall development of technology.

2 Relevant Research

In the context of the comprehensive development of China's comprehensive national strength, the notary system has played a key role in legal certification, financial risk prevention and control, and dispute resolution [1]. With the growth of market demand and the progress of information technology, the notarial archives industry is faced with the challenge of ensuring stable development and credit protection in the cross-chain public certificate mechanism.

The application of blockchain technology raises the issue of data silos, limiting the exchange of value between different blockchains. At present, notary mechanism in cross-chain technology has been paid more and more attention, but in the process of notary selection, it faces the problems of insufficient trust value calculation and unequal distribution of benefits. Guo Z proposed a node selection model based on verifiable random numbers and an improved EigenTrust algorithm [2] to optimize the fairness and trust of notary selection, thereby improving the reliability of cross-chain transactions.

Blockchain is used in many fields, but the exchange of value between different chains is limited. The existing cross-chain notary system is faced with problems such as assessing the risk of singleness and collusion. Chen proposed a scheme based on edge cloud storage to improve notary reliability and cross-chain transaction security [3].

In the field of P2P energy trading, the credibility and privacy protection issues of blockchain are particularly prominent. Wang proposed a dual-layer energy blockchain network, which significantly improves the credibility of P2P transactions through optimized cross-chain interoperability technology, ring mapping encryption algorithm and

consensus verification subgroup [4], and has been verified in the Ordos power market in China, demonstrating the effectiveness of this mechanism in improving transaction security.

In banking, notaries play an important role in ensuring legal certainty. BYP Munandar studied the legal status and liability of notary covering letter in bank credit behavior [5], pointed out that the covering letter failed to meet the requirements of Article 1868 of the Civil Code and Article 38 of the Notary Law for authentic documents, and discussed its impact on notary liability.

O Widiyastuti studied the role of notary public in handling the credit agreement and guaranteed loan default of BRI Bank Tegal City branch [6], emphasized the importance of notary public in formulating authentic documents with legal protection, and analyzed the role and effect of notary public in handling default by adopting social law methods.

In the field of agricultural engineering document management, L Shi proposed a new cross-chain mechanism combining notary mechanism and government supervision node, which significantly improved the credibility of cross-chain document certification and effectively solved the challenges of cross-chain document certification in the field of agricultural engineering [7].

S Liu proposed a master-slave chain based domain name resolution service architecture [8], which combined the multi-signature notarization scheme and MSBFT consensus algorithm to improve the efficiency of cross-chain communication. The performance of the model in the cross-chain notarization mechanism is verified by simulation experiments, and the actual effect of the improved cross-chain communication scheme is demonstrated.

In response to security authentication and cross-chain communication problems in the Internet of Things environment, S Shao proposes the blockchain cross-chain Communication mechanism (IBE-BCIoT) based on Identity encryption (IBE) [9], which solves related security authentication problems by electing proxy nodes and using public keys to securely communicate with cross-chain notary public.

Cao proposed a cross-chain data traceability mechanism, which realized cross-domain data traceability and trust enhancement by establishing global authorization chain and access chain within each trust domain [10], combined with notary group-based technology. The introduction of notary group election model based on reputation value effectively improves the credibility of notary group and meets the demand of cross-domain data traceability.

In the era of quantum computing, Z Wang proposed a cross-chain transaction model of quantum multi-signature notarization mechanism and asset quantum freezing algorithm to improve the security and efficiency of blockchain transactions [11]. This mechanism can effectively prevent forgery and denial, and track down malicious notaries, adapting to the new challenges brought by quantum computing.

3 Optimization of Cross-Chain Notary Mechanism Based on Trust Evaluation Model

3.1 Design and Implementation of Trust Evaluation Model

In the exploration of blockchain technology, trust evaluation models are vital, particularly for notary mechanisms in cross-chain transactions. Existing models like EigenTrust and PeerTrust have shown efficacy in certain scenarios but often struggle with the complexities of cross-chain transactions. To address these challenges, this paper introduces the Multidimensional Blockchain Cross-Chain Trust Model (MBCTT), an innovative approach designed to enhance trust evaluation in such environments.

The MBCTT model is crafted to address the unique characteristics of cross-chain transactions by incorporating a range of evaluation indicators to measure notary nodes' trustworthiness comprehensively. The model employs transaction satisfaction as a fundamental metric. The specific formula is shown in (1).

$$d(j, k) = \frac{tv_{df}(j, k)}{tv_{df}(j, k) + gb_{jm}(j, k)} \quad (1)$$

quantifies the success rate of notary nodes in transactions, where $d(j, k)$ represents the satisfaction level of node i with notary node j , $tv_{df}(j, k)$ denotes the count of successful transactions, and $gb_{jm}(j, k)$ indicates the number of failed transactions. This metric provides clear insight into notary nodes' performance and helps mitigate risks associated with individual node anomalies.

To overcome the limitations of single-dimensional evaluations, the MBCTT model incorporates the average trust rate. This metric aggregates evaluations from the main transaction node and other related nodes, offering a more comprehensive reflection of the notary node's overall performance. The calculation method is shown in (2).

$$avg_trust(k) = \frac{1}{2}(d(j, k) + d(l, k)) + \frac{\sum_m e(m, k)}{2o} \quad (2)$$

In this formula, $d(j, k)$ and $d(l, k)$ denote the satisfaction ratings of the main transaction node for the notary node, while $e(m, k)$ represents the evaluations from other nodes. This holistic evaluation approach helps counteract biases and prevents distortion from malicious notary nodes.

Furthermore, the MBCTT model integrates the influence of transaction amounts on trust value by employing reward and penalty factors. The reward factor adjusts based on transaction volume and the notary node’s average trust rate to avoid sudden spikes in trust values and ensure evaluation stability. Conversely, the penalty factor introduces an adjustment mechanism based on transaction amounts to deter fraudulent behavior. The formulas are represented by (3) and (4).

$$\theta(j) = \left(1 - \frac{u}{n}\right) \times g(j) \times \left(1 - f^{-n \cdot \text{pofz}}\right) \tag{3}$$

$$\varepsilon(j) = n(j) \times \left(1 - f^{-n \cdot \text{pofz}}\right) \tag{4}$$

The traditional notary system usually consists of two steps: first, the transaction initiator deposits funds into the notary node’s account on the source chain, and then the notary node transfers the amount to the receiving account on the target chain. In practical operation, the model supports transactions through a margin pool account and elects notaries based on trust values during the transaction request stage. After the transaction is completed, the model evaluates the notary’s behavior and updates the trust value table to ensure the smooth progress of the transaction.

In cross chain transactions of blockchain technology, the structural differences and inconsistent data definitions between different blockchains pose many challenges to asset circulation. When conducting cross chain transactions, the first step is to set pre transaction parameters to verify the validity of the transaction request. As shown in Table 1, the pre transaction parameter table includes the user accounts that initiate and receive transactions, transaction amounts, user signatures, and query numbers for the source and target chains. After receiving these parameters, the notary system will verify them, confirm the legality of the transaction, and then generate a formal transaction parameter table.

Table 1. Pre transaction Parameters

Number	Symbol	Variable name
1	Addj	Transaction initiator account
2	Addk	Transaction recipient account
3	Amount	Transaction amount
4	Sigj	Initiator’s signature
5	SrcQuery	Source chain query tag
6	DstQuery	Target chain query tag

The official trading parameter table has added more information compared to the pre trading parameter table, such as the notary’s account, the deposit pool’s account, the trading limit time, and the trading order number.

3.2 Performance Verification and Optimization Experiment

The introduction of the MBCTT trust evaluation model aims to improve the reliability of the notary mechanism and enhance the performance of cross-chain transactions. Experimental results demonstrate that the MBCTT model offers significant advantages over both the Peer Trust local trust model and the Eigen Trust global trust model under various testing scenarios.

In environments without malicious notary nodes, the transaction success rates across the models are comparable, with the MBCTT model performing similarly to Peer Trust and Eigen Trust. However, when evaluating trading time, the MBCTT model shows a slightly longer average processing time compared to Eigen Trust. This is attributed to Eigen Trust’s simpler computation and quicker convergence, which generally lead to faster transaction processing. In contrast, the MBCTT and Peer Trust models involve more complex evaluation parameters, resulting in marginally extended processing times. The following table shows the transaction success rate and transaction time data of each model under the condition of no malicious nodes, as shown in Table 2.

Table 2. Transaction success rate and transaction time data when there are no malicious nodes

Model	Transaction success rate (%)	Average trading time (seconds)
MBCTT	99.8	23.5
PeerTrust	99.8	24.0
EigenTrust	99.8	21.0

In the presence of malicious notary nodes, the MBCTT model significantly outperforms the PeerTrust and EigenTrust models. When the proportion of malicious nodes is 10%, the transaction success rate of the MBCTT model is 95.5%, significantly higher than the 91.0% and 90.0% of the PeerTrust and EigenTrust models. When the proportion of malicious nodes increases to 20%, the transaction success rate of the MBCTT model is 89.0%, which is still higher than the 84.0% and 82.5% of the PeerTrust and EigenTrust models. The following table shows the transaction success rate and transaction time data of each model under different proportions of malicious nodes, as shown in Table 3.

In the absence of malicious notary nodes, as the number of transactions increases, the cross chain transaction time of the MEMTEM model is slightly higher than that of the notary mechanism cross chain model, but the difference is not significant. This indicates that introducing a trust evaluation model will not incur significant additional overhead on transaction time. The specific experimental data shows that the time comparison between MEMTEM and notary mechanism cross chain model under different transaction quantities at 0 o’clock is shown in Fig. 1.

In terms of transaction time, with the increase of malicious notary nodes, both the notary mechanism model and MEMTEM’s transaction time have increased. This is due to the involvement of malicious nodes causing transaction timeouts and compensation operations, which increases processing time. However, the increase in transaction time of MEMTEM is significantly lower than that of the notary mechanism model, especially

Table 3. Transaction success rate and transaction time data for different proportions of malicious nodes

Proportion of malicious nodes (%)	Model	Transaction success rate (%)	Average trading time (seconds)
0	MBCTT	99.8	23.5
0	PeerTrust	99.8	24.0
0	EigenTrust	99.8	21.0
10	MBCTT	95.5	35.0
10	PeerTrust	91.0	38.0
10	EigenTrust	90.0	37.5
20	MBCTT	89.0	42.0
20	PeerTrust	84.0	45.0
20	EigenTrust	82.5	44.5

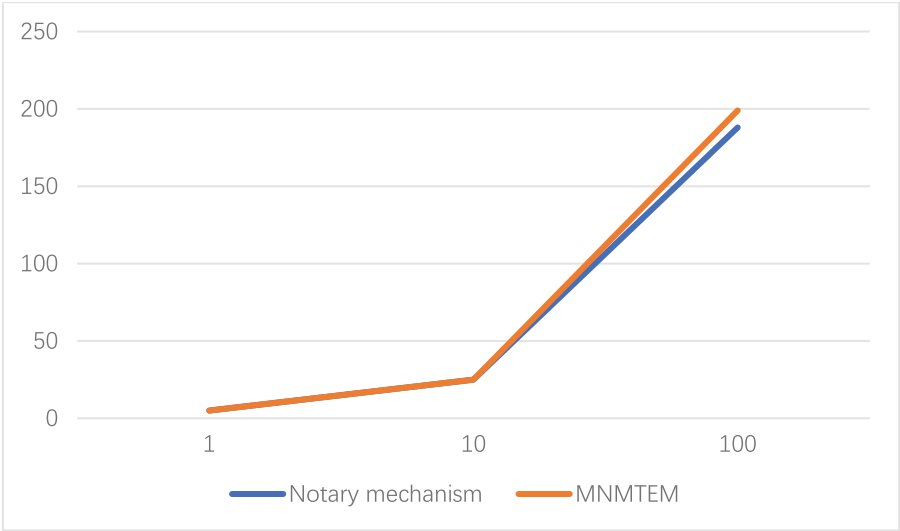


Fig. 1. Time comparison of the model under different transaction quantities at 0 o'clock

when the number of transactions is large, and the gap between the two is even more significant. This indicates that MEMTEM performs better in dealing with the additional time overhead caused by malicious nodes.

The MEMTEM model exhibits higher transaction success rates and better time performance in cross chain transaction scenarios, especially when facing malicious notary nodes. Compared with the notary mechanism model, MEMTEM can effectively improve the stability and efficiency of cross chain transactions. These experimental

results indicate that MEMTEM can significantly improve the performance of cross chain transactions and enhance the overall reliability of the system in practical applications.

4 Optimization of Notary Public Election Mechanism Based on Deep Q-Network

4.1 Design and Modeling of DQN Notary Public Election Algorithm

Reinforcement learning algorithms mainly include Monte Carlo reinforcement learning algorithm and time difference reinforcement learning algorithm. The Monte Carlo algorithm estimates the value function of a state through complete sampling, but in scenarios where the distribution of action rewards is particularly large, the variance of return values is large and the optimal strategy may not be found. In contrast, time difference reinforcement learning algorithms estimate the return value of the current state based on the parameters of the next state, with a smaller variance in the return value, while avoiding situations where the optimal strategy cannot be found.

Q-learning algorithm is a time difference method with different strategies, which guides learning through the Q-values corresponding to states and actions. The intelligent agent selects actions to execute based on a certain strategy in the current state, and updates the Q-table through state transitions and reward values. The Q-table contains Q-value information for state action, where the update formula for the Q-value function is shown in (5).

$$R_{u+1}(t, b) = (1 - \beta)R_u + \beta(s_u + \gamma nbyR_u(t, b)) \quad (5)$$

In this formula, β represents the learning rate, which controls the balance between the model's choice of results and experience; Gamma is a discount factor used to balance current returns with future long-term returns.

However, Q-learning algorithms face challenges when dealing with large-scale state spaces or action spaces, as Q-tables require storage of large amounts of data, slow computation speed. The DQN algorithm combines the advantages of reinforcement learning and neural networks, estimating Q-values through neural networks, thereby avoiding the computational and storage problems caused by Q-tables. Finally, by comparing the training effects of different discount factors, $\delta = 0.97$ was determined as the optimal discount factor selection, which can maintain good training stability and effectiveness while considering future returns.

4.2 Algorithm Performance Validation and Effect Evaluation

The experimental environment configuration of this study includes two parts of hardware and software settings: the virtual machine is equipped with 4G memory and Intel ® Core™ I7-10710U processor and 40GB hard drive, operating system is Ubuntu 16.04. Two blockchain networks based on Fabric 1.4 have been built on this virtual machine. The host part is configured with 16GB of memory and Intel ® Core™ I3-9100F processor, running Windows 10 operating system, used for deploying notary system. Throughout the development process, the blockchain system was implemented using Fabric 1.4 and

Docker, smart contracts were written in Go language, and the notary system and its trust evaluation model were implemented using Fabric sdk java. The development platform is Jupyter, and the training of deep learning models uses the TensorFlow framework with Python programming language.

In the design of the reward function, the aim is to balance the performance of cross chain transactions with the fairness of notary elections. The form of the reward function is as shown in formula (6).

$$S = s_0 + \theta U_{bmm-t} - \phi A_t \quad (6)$$

Among them, θ and ϕ are the weight coefficients of the sum of trust values and the variance of trust values, respectively, satisfying $\theta + \phi = 1$. After multiple rounds of experimental adjustments, $\theta = 0.55$ and $\phi = 0.45$ were determined as the optimal configurations to achieve the best balance between system performance and fairness.

The experimental design is based on the aforementioned model, using TensorFlow framework and Python to write DQN algorithm, and integrating it into a trust evaluation based notary cross chain system to simulate the actual transaction process for verification. The experiment simulates user transaction requests through Apache JMeter, and the notary system subsequently processes these requests and performs corresponding cross chain transaction operations.

The experiment is divided into three main parts: the first part tests the transaction success rate under different proportions of malicious notary nodes (0%, 10%, 20%, 30%, 40%), conducting 1000 and 5000 transactions respectively to compare the performance of notary mechanism cross chain model, MNMTEM, and MNMTEM-DQN. The second part evaluates the transaction time efficiency under different proportions of malicious notary nodes, conducting 100, 200, 300, and 400 transactions, and comparing the processing time of the three models. The third part analyzes the impact of the proportion of malicious nodes on the trust value of notaries, comparing the performance of MNMTEM and MNMTEM-DQN in trust value changes through 1000 and 5000 transactions.

The experimental results show that there are differences in the performance of different notary election models in cross chain transactions. As the number of malicious notary nodes increases, all testing models, including notary mechanism cross chain model, MEMTEM, and MNMTEM-DQN, have experienced a decrease in transaction success rates, as shown in Figs. 2 and 3. Although the success rate of MNMTEM-DQN has slightly decreased compared to MEMTEM, its transaction success rate is significantly higher than that of traditional notary mechanism models, especially when the number of transactions increases, this advantage becomes more significant, demonstrating the effectiveness of MNMTEM-DQN in malicious node environments.

In the comparison of transaction times, the results show that when there are no malicious nodes, the cross chain transaction time overhead of MNMTEM-DQN is higher than that of the notary mechanism and MEMTEM. This is mainly due to the additional time required by MNMTEM-DQN to handle trust evaluations and DQN notary elections. However, when malicious notary nodes appear, the time cost of MNMTEM-DQN gradually decreases compared to traditional notary mechanisms, indicating that MNMTEM-DQN improves the success rate of cross chain transactions and effectively reduces time costs. Although the time cost of MNMTEM-DQN is always higher than MEMTEM, this

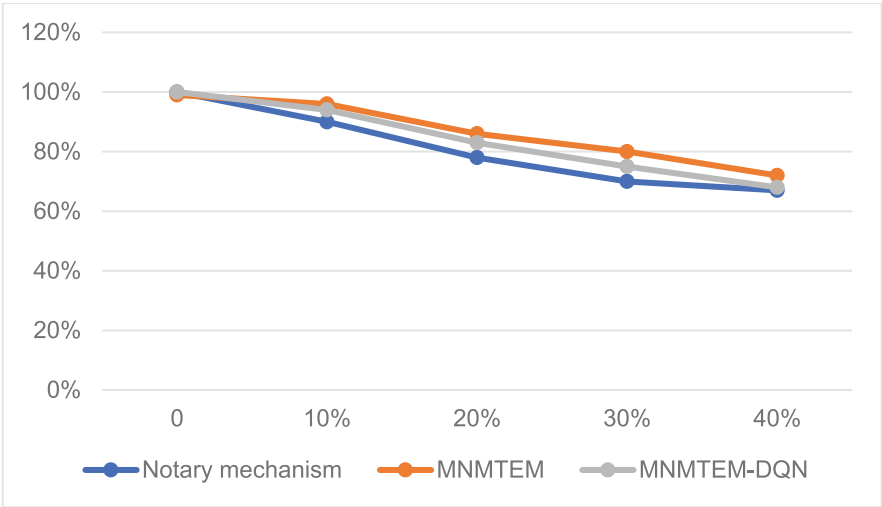


Fig. 2. Comparison data of transaction success rate when the number of transactions is 1000

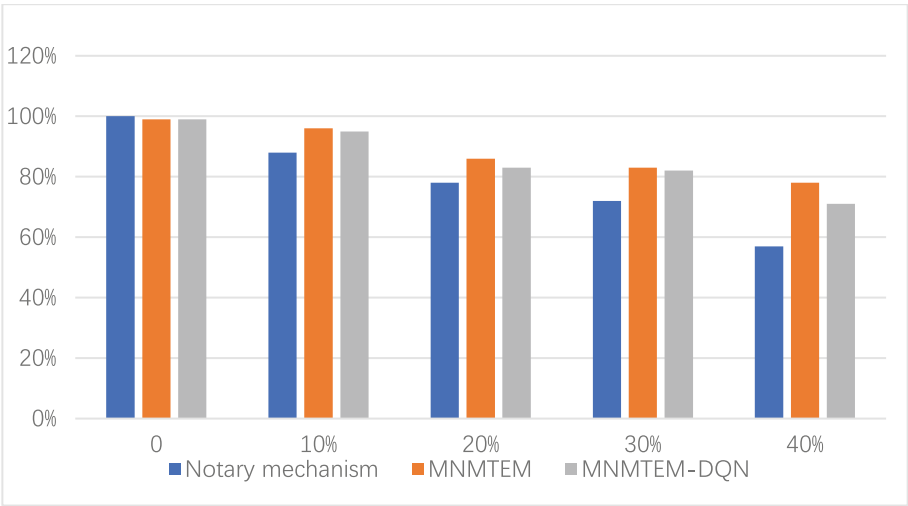


Fig. 3. Comparison data of transaction success rate when the number of transactions is 5000

can be attributed to MNMTEM-DQN sacrificing a portion of the transaction success rate in some cases to avoid centralization of notary elections.

The experimental results show that MNMTEM-DQN achieves a more balanced distribution of trust values for notaries after the transaction is completed than MEMTEM. MNMTEM-DQN effectively avoids the centralization of notary nodes and ensures the fairness of elections. However, in some cases, the trust values of certain notaries in MNMTEM-DQN are low, which reflects that in the pursuit of fair elections, the model

may sacrifice some transaction success rates, leading to an increase in the frequency of malicious nodes being selected.

5 Conclusion

This article systematically studies the integration of trust evaluation models and notary mechanism cross chain models. Firstly, it reveals the limitations of notary mechanism in cross chain transactions, especially its high dependence on notary reliability. In order to improve the reliability of notary nodes and reduce the repeated participation of malicious nodes, this paper proposes a new trust evaluation model - MBCTT, which comprehensively considers multiple factors such as penalty factors, reward factors, transaction amounts, and time to achieve a more comprehensive evaluation of node behavior. By combining the MBCTT model with the notary mechanism cross chain model, the MNMTEM model was formed, which effectively improves cross chain transaction performance and reduces the impact of malicious nodes through optimized transaction processes and protocols. At the same time, this article also introduces a notary election method based on DQN algorithm, which optimizes each link in the election process from the perspective of reinforcement learning to solve the problem of notary centralization. Although certain achievements have been made, further improvement in the verification of heterogeneous chains and in-depth exploration of the improvement of DQN algorithm in cross chain performance are still needed in future research.

References

1. Guo, Z., Hu, X.: Calculation and selection scheme of node reputation values for notary mechanism in cross-chain. *J. Supercomput.* **80**(12), 18177–18198 (2024). <https://doi.org/10.1007/s11227-024-06152-3>
2. Zhang, H.: Discussion on the path exploration of the innovation of notary archives management under the background of informatization. *Foreign Lang. Sci. Technol. J. Datab. (Abstract Edition) Econ. Manage.* (1), 69–72 (2022)
3. Chen, L., Chen, Y., Tan, C., et al.: Cross-chain asset trading scheme for notaries based on edge cloud storage. *J. Cloud Comp.* **13**(1) (2024). <https://doi.org/10.1186/s13677-024-00648-2>
4. Wang, L., Xie, Y., Zhang, D., et al.: Credible peer-to-peer trading with double-layer energy blockchain network in distributed electricity markets. *Electronics* (2021). <https://doi.org/10.3390/electronics10151815>
5. Munandar, B.Y.P.: Legal standing of notary covernote in making of banking credit deeds which result in criminal acts of corruption (case study of pangkal pinang state court decision no. 21/PID.SUS-TPK/2021/PN.PGP). *JISIP (Jurnal Ilmu Sosial dan Pendidikan)* (2023). <https://doi.org/10.58258/jisip.v7i1.4245>
6. Widiyastuti, O.: Notaries role analysis in implementation of credit agreements & defaults settlement with guaranteed liability. *Sultan Agung Notary Law Review* (2021). <https://doi.org/10.30659/SANLAR.V3I3.16328>
7. Shi, L., Zhou, Y., Wang, W., et al.: A cross-chain mechanism for agricultural engineering document management blockchain in the context of big data. *Big Data Research* **36** (2024). <https://doi.org/10.1016/j.bdr.2024.100459>

8. Liu, S., et al.: Domain name service mechanism based on master-slave Chain. *Intel. Auto. Soft Comp.* **32**(2), 951–962 (2022)
9. Shao, S., Chen, F., Xiao, X., et al.: IBE-BCIoT: an IBE based cross-chain communication mechanism of blockchain in IoT (2021). <https://doi.org/10.1007/s11280-021-00864-9>
10. Cao, L., Zhao, S., Gao, Z.S., et al.: Cross-chain data traceability mechanism for cross-domain access. *The J. Supercomp.* 1–18 (2022). <https://doi.org/10.1007/s11227-022-04793-w>
11. Wang, Z., Li, J., Chen, X.B., et al.: A secure cross-chain transaction model based on quantum multi-signature. *Quantum Information Processing* **21**(8) (2022). <https://doi.org/10.1007/s1128-022-03600-y>



Network Security Situation Automatic Prediction System Based on Artificial Intelligence

Wenyue Qi^(✉)

Queen Mary University of London, Mile End Road, London E1 4NS, UK
qi_wenyue@hotmail.com

Abstract. Network threats are an obstacle to the protection of network and information security, and the current prediction of network security situation needs to be further improved. This study adopts the method based on artificial intelligence, and focuses on comparing the performance of Bayesian classification algorithm, SVM and RNN in the prediction of network security situation. Bayesian classification algorithm is based on the principle of probability statistics and can classify and predict network security events. RNN has the ability to process sequential data and capture temporal relationships, which is suitable for time-dependent modeling of network security data. SVM can handle small samples and noisy data, and has high stability and reliability in network security prediction. The experimental results show that the Bayes classification algorithm has the advantage of shorter execution time, which is suitable for the network security prediction scenario with high real-time requirement. RNNs, however, perform best in terms of prediction accuracy, and are better able to capture the temporal evolution trends and complex patterns of cybersecurity events.

Keywords: Network Security · Situation Prediction · Artificial Intelligence · RNN

1 Introduction

Security threats such as cyber attacks, data leaks, and malicious behavior are constantly increasing, bringing huge risks and losses to individuals, organizations, and society. In order to effectively respond to these threats, it has become a vital task to predict and identify the network security situation in advance. The purpose of this article is to study the automatic prediction system of network security posture based on artificial intelligence, and to explore the comparison of the prediction performance of Bayesian Classifier, Support Vector Machine (SVM) and Recurrent Neural Network (RNN). By comparing the performance of the three algorithms in network security situation prediction, it can provide reference and guidance for building an efficient and accurate network security prediction system. This helps to detect and respond to cyber threats in a timely manner, and improves the security and stability of the network. In addition, this article

will also discuss the trade-off between the execution time of the algorithm and the prediction accuracy rate, and provide guidance for the selection of algorithms in practical applications.

This paper first introduces the background and significance of network security situation prediction, expounds the importance and value of this research method, then compares the performance of Bayesian classification algorithm and RNN in network security situation prediction, and finally summarizes the research results and puts forward the prospect of future research.

2 Related Work

Many scholars have studied network security situation. Zhang Jinlong proposed a sensor network security prediction method based on multi-source data binding fusion. This method uses convolutional self-coding network to represent the unified dimension of heterogeneous data of different modes, which can solve the problem of heterogeneity among sensors. Experiments showed that the proposed method was more robust than the Gram angular field algorithm integrated with convolutional neural network [1]. Ding Zhi proposed an improved whale optimization algorithm based on inertial chaos, designed an adaptive inertial weight position update mechanism, and balanced the global search ability and local development ability. Through experiments, he proved that the algorithm had better performance of jumping off local optimality and could improve search accuracy and convergence speed, and verified the feasibility and effectiveness of the algorithm through network security situation prediction [2]. According to the multi-factor characteristics of network security, Wan Bin analyzed the characteristics of log files, applied the Apriori algorithm, the association rule, to the direction of network security prediction, and found out the potential correlation between the log parameters when network security accidents occurred. He proposed a low-cost, large-capacity, high-efficiency network security prediction method model, and proposed a theoretical research basis for promoting the construction of information security system [3]. Qin Lina proposed a network security situation prediction method based on Kalman filter. The simulation results showed that the proposed model could accurately and effectively predict the network security situation, and had the advantage of high prediction accuracy, and had greater advantages compared with other methods [4]. Luo Hongfang proposed an optical network security situation prediction method based on Bayesian attack graph. He used vulnerability scanning software to scan the defects in the network, obtains initial data, establishes Bayesian attack graph based on relevant data, analyzed the uncertain parameters in the network, and calculated the prior probability and posterior probability of each node [5]. Wu D used binary semantic analysis to classify and predict network traffic data, which can effectively identify and predict various network attacks and improve the response ability of network security [6]. Wang H adopted a variety of machine learning algorithms to detect network intrusion and abnormal behavior [7]. Liaqat S focused on network security situation awareness and adopted hybrid detection to predict the development trend and possible threats of network security events [8]. Chen Z proposed an improved RBF neural network algorithm to model and classify network traffic data [9]. Yan W analyzed the historical network attack data through the Internet of Things and

predicted the future attack trend [10]. These methods provide great help for the prediction of network security situation. This paper will conduct in-depth research on the prediction of network security situation through artificial intelligence methods.

3 Method

3.1 Data Set Construction

This paper takes the prediction of network attack type as the specific goal of network security situation prediction, and selects the data source related to network attack. The data sources for this article are network traffic data, log files, intrusion detection system and firewall logs, security information and event management system data, malware samples, and other data related to network security. Based on the above data sources, this article collects and obtains data related to network attacks, including setting up network monitoring devices, configuring logging systems to integrate with security devices and systems, obtaining public data sets, or obtaining data from third-party security service providers. Cleaning the raw data collected, the original data is chaotic and contains a large number of invalid and redundant data, missing values and outliers. Data smoothing and feature scaling are used to process the data:

$$\text{Smooth value} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (1)$$

x_n represents the data point at the current moment, n is the size of the smoothing window, and feature scaling is used to unify the value ranges of different features:

$$X_{\text{scaled}} = \frac{X - X_{\text{mean}}}{X_{\text{std}}} \quad (2)$$

X is the original feature data, X_{mean} is the mean value of the feature, and X_{std} is the standard deviation of the feature. After processing, the data can be annotated and annotated to provide supervised learning training samples for the predictive model [11, 12], mark the attack types of network traffic data, and mark malicious behaviors in log files. When constructing the data set, the balance of samples should be considered to ensure that the number of samples in different categories is relatively balanced, so as to avoid the inaccuracy of prediction results caused by the bias of the model to a certain category.

3.2 Network Security Situation Awareness

Cybersecurity situational awareness is the process of collecting, analyzing, and interpreting cybersecurity related data to fully understand and identify current cybersecurity threats, attacks, and vulnerabilities, so as to take appropriate response and protection measures in a timely manner. To realize situation awareness, situation elements need to be extracted to describe and analyze the key features or indicators of network security situation. It extracts information from raw data to represent the status, trends, and

changes in cybersecurity. The selection and extraction of network security situation elements is to better understand and evaluate network security threats, attack behaviors and risks [13, 14]. The collected data is pre-processed, feature extracted, and analyzed to identify potential security threats, attacks, and unusual events. By analyzing and comparing data, cybersecurity situational awareness can identify known threats, malicious activities, and attacks. Cybersecurity situational awareness is not only about passively identifying threats, but also about taking timely response and protection measures to address threats, which can include automated response mechanisms, fixing vulnerabilities, isolating affected systems, updating security policies, and other measures to minimize potential losses and impacts. Network security situational awareness is an important part of ensuring network security. It helps organizations identify and respond to threats in a timely manner, and improves network security and resilience. Through continuous monitoring, analysis, and early warning [15], cybersecurity teams can better protect network assets and user data against increasingly complex and evolving cybersecurity threats.

3.3 Network Security Situation Automatic Prediction Based on Artificial Intelligence

Before situation prediction is realized, it is necessary to establish an evaluation system for security situation. Network security situation assessment system is a framework for comprehensive assessment and quantitative analysis of network security conditions, which includes a set of indicators and corresponding weights to measure various aspects of network security. Selecting network attack activity, security incident response ability, security vulnerability management, network device configuration security, and user security awareness and training from the indicators. The weight distribution is shown in Table 1:

Table 1. Weight distribution table

Indicator	Weight	Actual Value	Normalized Value	Weighted Score
Network Attack Activity	0.25	120	0.8	0.2
Security Incident Response Capability	0.2	90	0.6	0.12
Security Vulnerability Management	0.15	80	0.5	0.075
Network Device Configuration Security	0.15	95	0.63	0.0945
User Security Awareness and Training	0.1	70	0.46	0.046

In Table 1, each indicator contains the indicator name, weight, actual value, normalized value, and weighted score. The actual value is the specific observed value of the

index in a certain period of time, the normalized value is the result after normalization according to the actual value, and the weighted score is the result calculated according to the normalized value and weight.

The purpose of network security situation prediction is to discover and deal with potential network security threats in advance, and ensure the stable and secure operation of network systems. Through continuous analysis and optimization of the prediction model, combined with real-time network data and intelligence information, the prediction accuracy and response ability can be improved to effectively deal with increasingly complex and diversified network security threats. After the evaluation system is constructed, the current mainstream prediction algorithms include Bayesian Classifier, Support Vector Machine(SVM) and Recurrent Neural networks (RNN). The training and prediction process of Bayes classifier is relatively fast, especially suitable for large-scale data sets and real-time prediction scenarios. By maximizing the classification interval, SVM has good generalization ability, can handle small samples and noisy data, and has high stability and reliability in network security prediction. Recurrent neural networks transmit information by hiding the state and retain the historical context, which can remember and use the past information, and is suitable for modeling and predicting the time dependence of network security events. This paper will test the predictive ability of these three algorithms.

4 Results and Discussion

Three prediction methods were mentioned above. In this chapter, the performance of these three methods will be compared. First, the data processed in this paper will be made into a data set, and the convergence of the three algorithms will be tested respectively (Fig. 1).

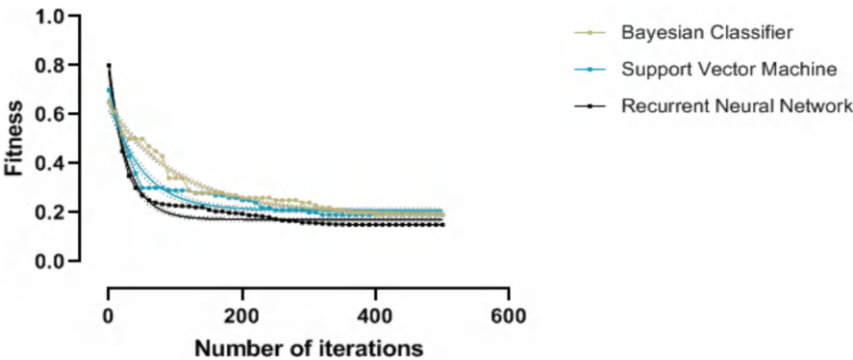


Fig. 1. Algorithm convergence

In the convergence test of the above algorithms, the fitness value of Bayes classification algorithm reaches a small value of 0.29 after 90 iterations, and a minimum value of 0.19 after 400 iterations. The SVM algorithm reaches the minimum value 0.30 after 50 iterations and the minimum value 0.19 at the 320 iteration. The recurrent neural network

reaches a smaller value of 0.25 after 60 iterations and a minimum value of 0.15 at 360 iterations. Among the three algorithms, support vector machine reaches the minimum fitness value after fewer iterations, and its algorithm convergence performance is the best. In order to better compare the accuracy of the three algorithms in situation prediction, this paper will conduct performance tests on them, and the test indicators are execution time, prediction accuracy and comparison between predicted value and actual value.

4.1 Execution Time

Execution time can be used to evaluate the efficiency and computational speed of the algorithm, and real-time is very important in cybersecurity situation prediction, because responding to and handling security incidents in a timely manner can reduce potential threats and damage. The shorter execution time means that the model is able to process input data faster and generate predictive results, which improves response speed and efficiency. Figure 2 shows the results of the execution time comparison:

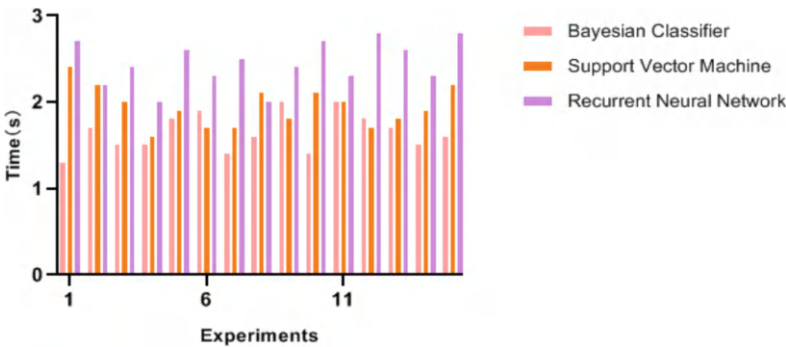


Fig. 2. Execution time

The test results of execution time show that the maximum time of Bayes classification algorithm is 2 s and the minimum time is 1.3 s, the maximum time of support vector machine algorithm is 2.4 s and the minimum time of 1.6 s, and the maximum time of RNN algorithm is 2.8 s and the minimum time of 2.0 s. Bayes has the advantage in terms of execution time, it can perform the prediction in a shorter time, because the training and prediction process of Bayes classifier is relatively fast, especially for large-scale data sets and real-time prediction scenarios.

4.2 Prediction Accuracy

Prediction accuracy measures the model’s ability to correctly classify samples. In network security situation prediction, accuracy can tell us how accurate the model is in predicting different types of security events or attacks. The high prediction accuracy means that the model can accurately classify normal and abnormal network traffic or

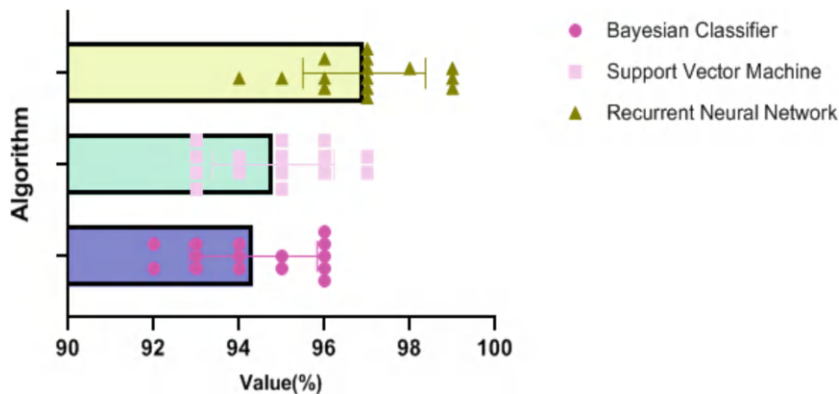


Fig. 3. Prediction accuracy

behavior, thus improving the detection and identification of security events. Figure 3 shows the experimental results:

In the test of prediction accuracy in Fig. 3, RNN is the algorithm with the best performance, and its prediction accuracy can reach 99% at the highest, while SVM algorithm can reach 97% at the highest, and Bayesian classification algorithm can reach 96% at the highest. High prediction accuracy can improve the credibility and reliability of the model. When the accuracy of automatic prediction of network security situation is high, we can rely more confidently on the prediction results of the model and take corresponding security measures based on these results.

4.3 Deviation Between Predicted Value and Actual Value

Biases can provide clues about the confidence of the predicted results, and smaller biases may indicate that the model is highly predictive for different categories or events, thereby enhancing the confidence of the predicted results. Conversely, large deviations may reduce the confidence of the predicted results and require further review and adjustment of the model. Figure 4 shows the deviation result:

Deviation experiment results show that the maximum deviation between Bayesian classification algorithm and the actual value is 0.08, the maximum deviation between SVM algorithm and the actual value is 0.06, and the maximum deviation between RNN neural network is 0.04. This is because RNNs are suitable for processing sequence data, while network security situation prediction usually involves time series data, RNNs have memory and context awareness capabilities, and can capture timing information and association relationships in input sequences.

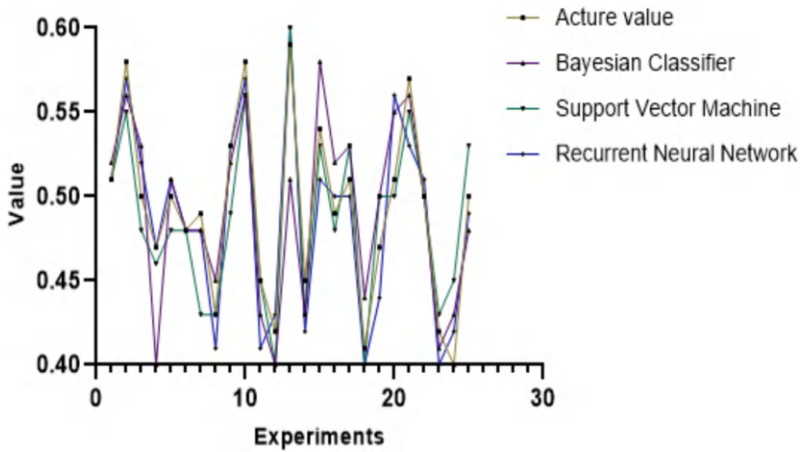


Fig. 4. Deviation

5 Conclusion

Through the comparative testing of several indicators, it is found that different algorithms show their own characteristics in terms of prediction performance and execution efficiency. In this study, we compared the performance of Bayesian classification algorithms and RNN and SVM in network security posture prediction. We have observed that Bayesian classification algorithms have obvious advantages in execution time. Since the Bayesian algorithm is based on the principle of probabilistic statistics, its computational complexity is low and it can complete the prediction task in a shorter period of time. At the same time, it was found that RNN performed best in terms of prediction accuracy. RNN has the ability to process sequence data and capture timing relationships, and can better predict the time evolution trend of network security events, which makes RNN have higher accuracy and reliability in network security situation prediction.

References

1. Jinlong, Z., Guotong, S.: Sensor Network security prediction Model based on multi-source data binding fusion. *Comm. Power Technol.* **40**(16), 178–181 (2023)
2. Zhi, D., Yu, X.: Improved whale optimization algorithm and network security prediction based on inertial chaos. *J. Xinxiang Univ.* **39**(6), 44–49+62 (2024)
3. Bin, W., Ming, X.: A network security prediction method based on Apriori algorithm. *Elect. Power Info. Comm. Technol.* **17**(1), 133–138 (2019)
4. Lina, Q.: Network security prediction based on kalman filter. *J. Xiangnan Univ.* **40**(2), 22–25 (2019)
5. Luo, H., Wang, C.: Research on optical network security situation prediction based on Bayesian attack graph. *Laser Journal* **44**(8), 134–138 (2023)
6. Wu, D.: A network security posture assessment model based on binary semantic analysis. *Soft. Comput.* **26**(20), 10599–10606 (2022)
7. Wang, H., Zhao, D., Li, X.: Research on network security situation assessment and forecasting technology. *J. Web Eng.* **19**(7–8), 1239–1266 (2020)

8. Liaqat, S., Dashtipour, K., Arshad, K., et al.: A hybrid posture detection framework: integrating machine learning and deep neural networks. *IEEE Sens. J.* **21**(7), 9515–9522 (2021)
9. Chen, Z.: Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. *J. Computat. Cognit. Eng.* **1**(3), 103–108 (2022)
10. Yan, W., Qiao, L., Krishnapriya, S., et al.: Research on prediction of school computer network security situation based on IOT. *Int. J. Sys. Assur. Eng. Manage.* **13**(Suppl 1), 488–495 (2022)
11. David, A.O., Oluwasola, O.O.: Zero day attack prediction with parameter setting using Bi direction recurrent neural network in cyber security. *Int. J. Comp. Sci. Info. Sec. (IJCSIS)* **18**(3), 111–118 (2020)
12. Li, J., Wu, Y., Li, Y., et al.: A network security prediction method based on attack defense tree. *J. Nanoelectron. Optoelectron.* **18**(3), 357–366 (2023)
13. Lu, Y., Kuang, Y., Yang, Q.: Intelligent prediction of network security situations based on deep reinforcement learning algorithm. *Scala. Comp. Pract. Exper.* **25**(1), 147–155 (2024)
14. Dong, R.H., Shu, C., Zhang, Q.Y., et al.: Security situation prediction method for industrial control network based on adaptive Grey Verhulst model and GRU network. *Int. J. Netw. Sec.* **24**(1), 49–61 (2022)
15. Du, J., Yuan, F., Ding, L., et al.: Research on threat information network based on link prediction. *Int. J. Digi. Crime and Forensics (IJDCF)* **13**(2), 94–102 (2021)



Research on Optimization Algorithm of Multi-agent System

Jieru Wang^(✉)

Engineering Electrical and Electronic Engineering, Taylor's University, Subang Jaya 47500,
Selangor, Malaysia
wangjieru2000@gmail.com

Abstract. Multi-agent system technology has been widely used in power system regulation, warehousing and logistics, etc. However, the accurate evaluation and deep optimization of the performance of such systems is still a frontier issue to be solved, and it is still mainly in the experimental and exploration stage. In this paper, a new attribute-based multi-agent design method is proposed, which is combined with performance analysis to carry out performance evaluation in the system design process. The design method takes the necessary attributes of the system as the core, adopts the top-down top-level design logic, starts from the functional requirements at the macro level, and refines step by step to the specific realization at the micro level, so as to ensure that the system design is closely developed around the actual needs. We have incorporated the “Off-Policy” evaluation method into this system and integrated it into the model verification step of the system design step. This evaluation method has unique advantages in the field of reinforcement learning, which can evaluate and compare the potential effects of other strategies while keeping the current operational strategies unchanged, and provide strong support for coping with the complexity and dynamic challenges in multi-agent systems. During the design process, we will carry out continuous evaluation and verification of the system: if the system can meet the preset attribute requirements, the design will enter the subsequent implementation phase to continue to promote; Conversely, if the system fails to meet the preset attribute criteria, an iterative optimization process is initiated to continuously adjust and refine the design until all attribute requirements are met.

Keywords: Multi-Agent · Markov Chain · Off-Policy Evaluation

1 Introduction

In today's ever-changing era of science and technology, multi-agent system has become an effective tool to solve complex problems, in many fields, including but not limited to power system scheduling, warehousing logistics management and large-scale robot collaborative work scenarios, the application of multi-agent system increasingly highlights its unique advantages [1–3]. However, at the beginning of the design, it has been challenging for scholars to know exactly how to take full account of and satisfy the expected attribute requirements of a system and also appraise its performance to enable optimization.

The method of design of multi-agent systems driven by demand for attribute put forward in this paper is a novel thought to respond to the challenge. This approach stipulates that the design process should be oriented towards requirements of attributes, i.e. it should begin from top-level architecture and gradually refine downwards each layer until reaching underlying modules so as to ensure that all design elements and decisions made can be directly traceable back to serving core attributes needed by the system. This model helps in creating multi-agent solutions which are more customised to real needs and hence more specific and practicable.

Meanwhile, the Off-Policy evaluation technique was introduced to measure and enhance the performance of multi-agent systems. Off-Policy evaluation is an advanced reinforcement learning assessment method which enables the comparison and evaluation of different potential strategies besides the current execution strategy. This technique proves to be particularly advantageous when dealing with non-linear, dynamic change as well as randomization issues that are typical in multi-agent systems. Consequently, we also look into applying the Off-Policy evaluation method for performance assessment on macro-enhanced Markov model (MHA-MDP). In MHA-MDP, the time evolution characteristics of the system and the probability distribution of various possible state transitions are reflected, which enables us to accurately capture and quantify the performance indicators by simulating and analyzing the behavior response of the system over multiple time steps.

In specific operation, we use two statistical methods, common importance sampling (CIS) and weighted importance sampling (WIS), to reveal and evaluate the performance of multi-agent systems at different time steps in the process of iterative calculation. These two sampling methods can effectively overcome the sampling bias problem in Markov decision process, so as to provide more accurate and reliable system performance evaluation results.

In summary, this paper organically integrates the Off-Policy evaluation method into the attribute-based demand-driven multi-agent system design process to achieve in-depth analysis and real-time optimization of system performance in the design stage. This innovation is expected to promote the dual progress of multi-agent system design theory and practice, especially in complex multi-agent application scenarios such as multi-robot collaborative operation, and bring substantial breakthroughs in improving system efficiency, stability and robustness. In the future, with the further promotion and application of this method, we have reason to expect to see more multi-agent systems with excellent performance play a key role in all walks of life, leading a new round of technological innovation.

2 Related Work

In the real world, there are a large number of multi-agent cooperative decision-making tasks, each agent playing a different role to maximize the cumulative benefits of the entire team. Among them, how to make the right behavior decision for each agent in the process of cooperation is an urgent problem to be solved. The Multi-Agent Reinforcement Learning (MARL) method [4] can realize complex collaborative control of multiple agents, and has been widely used in tasks such as multi-robot control and resource

allocation [5]. Multi-agent reinforcement learning is a method based on reinforcement learning and deep learning. Reinforcement learning [6] Inspired by behaviorism theory in psychology, the agent learns the optimal strategy by constantly interacting with the environment to obtain reward feedback. That is, using a trial-and-error mechanism, by constantly adjusting the actions chosen by the agent, the intelligence can learn how to obtain the maximum benefit. Due to the small state and action space, the tasks handled by traditional reinforcement learning methods can usually be represented in tabular form, but when the input data is pictures and other data, the dimensionality is too large to extract features effectively. As an effective feature extraction technology, deep learning is widely used in image, natural language and other fields, so deep learning can make up for the shortcomings of reinforcement learning. Now, the combination of reinforcement learning and deep learning is everywhere, in the field of natural language processing (NLP) can be used to recommend ads; In the field of vision, it can handle image repair, target tracking and other tasks; In the financial field, quantitative trading, asset management, etc.; At the same time, the success of AlphaGo [7] and AlphaGo Zero [8] developed by DeepMind has pushed the research on reinforcement learning to another climax. Based on the success of deep reinforcement learning, it has promoted its development in the field of multi-agent. Multi-agent reinforcement learning usually adopts end-to-end training, so how to choose a suitable training framework is an urgent problem to be solved. Because the environment of the agent is dynamically changing, when the agent interacts with the environment, if other agents are regarded as part of the environment, the strategies of other agents are constantly changing, resulting in the instability of the environment. If all agents are simply trained as one agent, in this case, as the number of agents increases, the state and action space will increase exponentially, resulting in a dimensional explosion, and the agent may not be able to learn a good strategy. In order to address the above problems, many existing methods adopt the Training framework of Centralized Training with Decentralized Execution (CTDE) [9–11]. The whole bureau information is used in the training process. This problem can be solved to a certain extent when the agent selects actions only according to its local observations.

At the same time, aiming at the problem of partial observation of agents in the environment, scholars have developed various communication based multi-agent reinforcement learning algorithms to strengthen the cooperation of agents. With the development of attention mechanism and graph neural network, these techniques are also widely used in the field of multi-agent [12].

3 Theoretical Basis

3.1 Markov Chain

1) *Brief introduction of principle.*

$X_1, X_2, X_3 \dots$ is a Markov chain that describes a sequence of states, each of which depends on a finite number of previous states [13–15]. A Markov chain is a sequence of random variables with Markov properties. The range of these variables, the set of all their possible values, is called the “state space”, and the value of X_n is the state at time n .

If the conditional probability distribution of X_{a+1} for the past state is only a function of X , then there is:

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \quad (1)$$

where x is some state in the process. The above identity can be regarded as a Markov property.

2) *Related concepts and properties*

3) *Transition probability matrix*

Definition: Suppose that the discrete state of the system is $X_1, X_2, X_3 \dots$. The state space formed is E , A_j^k indicates that the system is transferred to the state X_j for the k time, and A_i^{k-1} indicates that the system is in the state X_i , before the transfer, then it is called $p_{ij}^k = p\left(\frac{A_j^k}{A_i^{k-1}}\right)$ is the transition probability of the system to the state X_j at the k th time, and the data matrix composed of P_{ij} is the state transition probability matrix of the system. The transition probability p_{ij}^k has the following property. First, p_{ij}^k only depends on the initial state X_i and the number of transition steps k , but not on the subsequent states. Secondly, satisfy

$$p_{ij}^k \geq 0, \forall i, j \in E \quad (2)$$

$$\sum_{j \in E} p_{ij}^k = 1, \forall i \in E \quad (3)$$

b) *Reducibility*

Markov chain is represented by a conditional distribution $P(X_{n+1} | X_n)$ this is called is in the process of random transition probability. This is sometimes called the “one-step transition probability”. The transition probabilities for two, three, and more steps can be obtained from the one-step transition probabilities as follows.

$$P(X_{n+2} | X_n) = \int P(X_{n+2}, X_{n+1} | X_n) dX_{n+1} \quad (4)$$

$$P(X_{n+3} | X_n) = \int P(X_{n+3} | X_{n+2}) \int P(X_{n+2} | X_{n+1}) P(X_{n+1} | X_n) dX_{n+1} dX_{n+2} \quad (5)$$

The above formula can be generalized to any future time $n + k$ by multiplying the transition probabilities and integrating $k-1$ times.

c) *Periodicity*

Let $\{X_m, m \in E\}$ be a homogeneous Markov chain with state space E . For a state X , if the state set $\{m : m \geq 1, p_x^m > 0\}$ is nonempty, then the greatest common divisor L of the set is the period of state X , denoted as $d(x)$. A state m is said to be periodic if $L > 1$, and a state X is said to be aperiodic if $L = 1$.

d) *Decomposition of the state space*

Let $\{X_m, m \in E\}$ be a homogeneous Markov chain with state space E and C any subset of E . C is said to be closed if no state outside C can be reached from any state in C . For any $i \in C, k \notin C$ has $p_{ik} = 0$, then C is irreducible, and is called an irreducible Markov chain if there is no Markov chain of closed sets except the entire state space.

e)Universality

Let $\{X_m, m \in E\}$ as the homogeneous markov chain, the state space for the E, for any $i, j \in E$, if there is not dependent on constant π_j , I make $\lim_{n \rightarrow \infty} p_{ij}^{(m)} = \pi_j$, then the markov chain is universal. For a Markov chain with ergodicity, no matter from which state of the system, when the number of transition steps m is sufficiently large, the probability of the system transitioning to state j is approximately equal to π_j . In other words, the system reaches a stationary state after a period of time.

A Markov chain with stationary distribution satisfies the following condition: Let $\{X_m, m \in E\}$ be a homogeneous Markov chain if there exists a set of real numbers π_j such that:

$$\pi_j \geq 0, j \in E \quad (6)$$

$$\sum_{j \in E} \pi_j = 1 \quad (7)$$

$$\pi_j = \sum_{i \in E} \pi_i p_{ij}, j \in E \quad (8)$$

In general, the stationary distribution of a homogeneous Markov chain is not unique, but its stationary state exists and is unique if the following conditions are satisfied. A homogeneous Markov chain is said to be universal if there exists a positive integer m such that, for any $i, j \in E$, the m -step transition probability is greater than 0. In this process, the steady-state probability π_j meets π_j stability $\sum_{i \in E} \pi_i p_{ij}, j \in E$. It can be proved that its interpretation is unique and exists, π_j meets the probability distribution:

$$\pi_j > 0, \sum_{j=1}^M \pi_j = 1 \quad (9)$$

$$\pi^* P = \pi^* \quad (10)$$

Here, π^* is the vector formed by the steady-state probabilities $\pi_j, j \in E$, and P is the transition probability formed by the transition probabilities rate matrix P_{ij} .

3.2 Probabilistic Temporal Logic

Probabilistic Temporal Logic (PCTL) is based on the concept of computation tree, which can capture or express the distribution of system information under the influence of time and probability, and is very suitable for the application of multi-agent systems. The root of the PCTL represents the initial state, each subsequent node represents the state that may be reached after one step, and edges connect these states to form multiple paths, the state paths that the system may execute. In each path, the order of the nodes represents the temporal evolution of the system execution state. In DTMC, each time step is fixed, while in CTMC, the time step is exponentially distributed.

In addition to the time factor, PCTL extends Computation Tree Logic (CTL) by introducing a probability factor. We can add a probability factor of the system from the previous state to the next state to each time step to describe the probability of a state

transition in a multi-agent system. Here we present an example to describe the essence and constraints of a robot in terms of PTL: First, we define a set of atomic propositions representing the state of the robot and the properties of the environment. Set denotes the proposition that the picking robot is at position I and is true at time t . Denote that the statement representing the delivery robot at position I is true at time t . Denote the proposition that the picking robot is idle at time t .

4 Model Design

4.1 Optimize the Process

In the process of developing a multi-agent system, the developer first defines the system requirements by elaborating on a set of desired properties. These properties include not only the ultimate mission goal that the system should achieve, but also properties such as stability and adaptability of the system throughout its operation, such as maintaining a minimum number of working robots. And the recovery time limit to ensure the security of information transmission in the face of external interference. Then, in order to construct the system model that meets these requirements, the macroscopic Markov model is used to describe the behavior and evolution law of the system, and the rigorous model checking and iterative optimization are carried out with the help of Probabilistic Temporal Logic (PCTL) to ensure that the model can theoretically reflect all the expected properties.

However, in this stage, there may be some uncertain numerical parameters in the model, which stems from problems such as interval uncertainty of the function solution or individual agents may get trapped in local optimal solutions. In order to solve these problems, model checking and improvement is divided into three steps. Firstly, the established model is simulated and analyzed, and the Off-Policy evaluation method is used to verify whether the performance of the system in the real situation can meet the preset property requirements. If the system behavior is found to be inconsistent with the prediction of the specification model, the developer needs to repeat the design process, adjust and improve the specification model, and make corresponding corrections to the low-level implementation of the system to ensure consistency at all levels from high-level requirements to concrete implementation.

Finally, after the above iterative design and verification steps, the multi-agent system design that meets all the preset requirements is successfully completed, which ensures the consistency and effectiveness of the system design and the actual operation effect. The model flow is shown in Fig. 1.

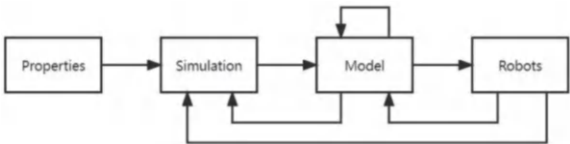


Fig. 1. Model flow

4.2 Collaborative Decision Making Based on Markov Decision and Probabilistic Temporal Logic

In this model, we combine Markov Decision Process (MDP) with Probabilistic Temporal Logic (PTL) to construct a collaborative decision-making framework for multi-agent systems. In particular:

1. Macro-enhanced Markov Decision Process (MA-MDP):

The system consists of multiple agents, each with its own action space and observation space, and all agents share a global state space S .

The agent selects actions according to its own strategy, transitions states according to a certain state transition probability matrix, and receives an immediate reward from the environment.

2. Probabilistic Temporal Logic (PTL) Goal Statement:

We exploit PTL to formally describe long-term behavioral specifications or performance metrics that a multi-agent system needs to satisfy. For example, “with 90% probability, the swarm of agents will reach a predetermined set of states in the next 10 steps”, or “For 5 consecutive cycles, at least one agent will remain inside the safe zone”, etc.

3. Off-policy evaluation and optimization:

The Off-Policy reinforcement learning algorithm is used to evaluate and optimize the policy of each agent, so that the behavior of the whole system meets the predefined PTL specification as much as possible.

In the evaluation process, common Importance Sampling (CIS) or weighted Importance sampling (WIS) methods are combined to correct the bias caused by Off-Policy and accurately calculate the probability of meeting the PTL specification.

4. Model solving and verification:

To address MA-MDP, a reinforcement learning system is created for agents that can adhere to the PTL specification during optimization of the overall cumulative reward.

In complex multi-agent situations, the proposed approach allows for productive collaborative decision-making through systemized analysis supported by simulations as well as theory verification ensuring efficient fulfilment of temporal logic constraints within a system.

5 Experiment

5.1 Evaluation Method

Methods for evaluating policies that are off-policy are crucial in reinforcement learning for multi-agent systems because they assess and improve system performance. The special thing about this method is its capability to appraise the behavior policy of interest while different from the currently running one. Such an ability is specifically important with respect to multi-agent systems which are characterized by their high complexity, uncertainty including non-linear dynamics and stochastic interactions as well as environmental changes among others.

The macroscopically augmented Markov models (MHA-MDP) are made for considering the general system evolution over time through accounting for the inter-agent interactions' effects. It comprises numerous entities' state transition matrices where each state does not only mirror one agent's status but also embraces the joint distribution of a full system state space.

Common Importance Sampling (CIS) and Weighted Importance Sampling (WIS) are two commonly used techniques when using Off-Policy evaluation. This is used to correct the bias caused by sampling from the action policy rather than the target policy. The key idea of these methods is to assign weights to each sampling sequence so that it can represent the expected value under the target policy.

The basic principles of CIS and WIS can be formally expressed as follows.

Assume that behavior strategy $\pi_b(a_t|s_t)$ determines the actual action is a probability, and target strategy $\pi_e(a_t|s_t)$ we want to evaluate strategies, the Importance weights (Importance Weight, IW) is defined as:

$$w_t = \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)} \quad (11)$$

Common Importance sampling (CIS) typically applies these weights at a single time step to adjust the cumulative Return:

$$G_t^{CIS} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \cdot w_{t+k} \quad (12)$$

Weighted Importance sampling (WIS) applies the product of weights over an episode to correct the expected value of the entire sequence:

$$G_t^{WIS} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \cdot \left(\prod_{i=t}^{t+k} w_i \right) \quad (13)$$

where γ is the discount factor.

Through this modification, the Value Function or Action-Value Function of the target policy at each time step can be estimated more accurately, and then the system performance index can be optimized.

5.2 Evaluation of Results

5.2.1 Comparative Model Selection

Multi-Agent Deep Deterministic Policy Gradient (MADDPG) is a reinforcement learning algorithm for multi-agent systems. It is an extension of DDPG (Deep Deterministic Policy Gradient) algorithm. MADDPG aims to solve the problem of cooperation and competition in multi-agent environment, so that each agent can learn an effective policy in a complex multi-agent environment.

DDPG (Deep Deterministic Policy Gradient) is an algorithm that combines deep learning and reinforcement learning techniques to solve control problems in continuous action Spaces. It is an actor-critic approach that utilizes deep neural networks to represent the policy (actor) and the value function (critic). The design of DDPG is inspired by

Table 1. Comparison results

Moels	Average waiting time	Task completion rate
DDPG	3 min	95%
MADDPG	4 min	93%
Our	2 min	97%

the Deterministic Policy Gradient (DPG) algorithm and the Deep Q-Network (DQN). It consists of the following key components:

The comparison results are shown in Table 1. When comparing the performance of DDPG, MADDPG and our own model on specific tasks, we find that the customized method has the best performance in average waiting time (2 min) and task completion rate (97%), showing the highest efficiency and effect. DDPG offered a balanced choice, coming in second with a 3-min wait time and a 95% task completion rate, while MADDPG performed worst when dealing with the complexity of multi-agent interactions with a 4-min wait time and a 93% task completion rate. This shows that in specific application scenarios, selecting or optimizing algorithms to adapt to specific needs can significantly improve the efficiency and success rate of task processing.

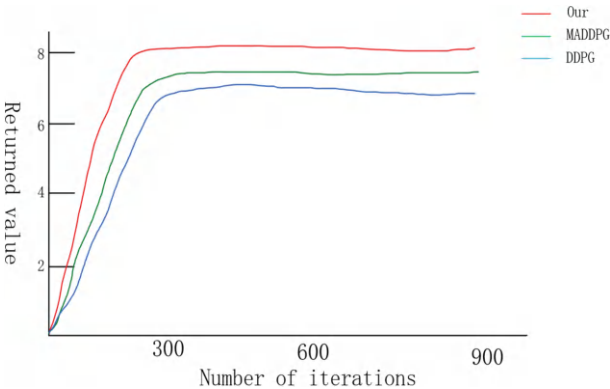


Fig. 2. The average return of each algorithm iteration

The Fig. 2 above shows that over multiple iterations, our method significantly outperforms MADDPG and DDPG methods in terms of performance metrics, indicating that it not only learns faster but also achieves higher stability. Although the MADDPG method started slowly, its performance improved steadily with the increase of iterations. In contrast, the DDPG approach, while rapidly improving performance in early iterations, quickly slows down and eventually reaches a relatively low level of performance. This suggests that our approach may be a better choice when fast and efficient optimization is required.

6 Summary and Prospect

In this paper, an innovative attribute-based multi-agent system design method is proposed, and combined with performance analysis technology, real-time performance evaluation is realized in the design process. The method takes various attributes of the actual requirements of the system as the core, and adopts top-down design logic to ensure that the design work is always closely related to the real scene. In this design framework, the “off-strategy” evaluation method in the field of reinforcement learning is introduced for the model verification stage, which can effectively evaluate and compare the possible effects of other strategies without changing the current operation strategy, and help to deal with the complexity and dynamic problems faced by multi-agent systems. During the design process, the system will undergo continuous evaluation and verification, and only when the system meets all preset attribute requirements will it enter the implementation phase, otherwise the design will be continuously improved through an iterative optimization process.

In the future, with the in-depth application of multi-agent systems in more industries and fields, the demand for performance evaluation and optimization will be more urgent. Attribute-based design method and integrated reinforcement learning evaluation technology are expected to be effective ways to solve this problem, and play a key role in multiple application scenarios such as power system regulation, warehousing and logistics. With the further research, this design method is expected to improve the stability and adaptability of multi-agent systems, and promote their realization of more efficient and accurate task execution capabilities. At the same time, the theoretical research and practical exploration of multi-agent systems will further enrich and improve the relevant evaluation system, and promote the rapid development of intelligent system science.

References

1. Ren, F., et al.: Conceptual design of a multi-agent system for interconnected power systems restoration. *IEEE Trans. Power Sys.* **27**(2), 732–740 (2012)
2. Draganjac, I., et al.: Decentralized control of multi-AGV systems in autonomous warehousing applications. *IEEE Trans. Auto. Sci. Eng.* **13**(4), 1433–1447 (2016)
3. Jabeur, N., et al.: Toward leveraging smart logistics collaboration with a multi-agent system based solution. *Procedia Computer Science* **109**, 672–679 (2017)
4. Busoniu, L., Babuska, R., De Schutter, B.: A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Sys. Man, and Cybernetics, Part C (Applications and Reviews)* **38**(2), 156–172 (2008)
5. Cui, J., Liu, Y., Nallanathan, A.: Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Trans. Wireless Commun.* **19**(2), 729–743 (2019)
6. Sutton, R., Barto, A.: *Reinforcement Learning: an Introduction*. MIT Press (1998)
7. Silver, D., Huang, A., Maddison, C.J., et al.: Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587), 484–489 (2016)
8. Silver, D., Schrittwieser, J., Simonyan, K., et al.: Mastering the game of go without human knowledge. *Nature* **550**(7676), 354–359 (2017)
9. Kraemer, L., Banerjee, B.: Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing* **190**, 82–94 (2016)

10. Xie, J., Liu, C.-C.: Multi-agent systems and their applications. *J. Int. Council on Electr. Eng.* **7**(1), 188–197 (2017)
11. Qin, J., et al.: Recent advances in consensus of multi-agent systems: a brief survey. *IEEE Trans. Indus. Electr.* **64**(6), 4972–4983 (2016)
12. Byrski, A., et al.: Evolutionary multi-agent systems. *The Knowl. Eng. Rev.* **30**(2), 171–186 (2015)
13. Norris, J.R.: *Markov chains*. No. 2. Cambridge University Press (1998)
14. Hu, Z., Xiaowu, M.: Impulsive consensus of stochastic multi-agent systems under semi-Markovian switching topologies and application. *Automatica* **150**, 110871 (2023)
15. Du, Y., Wang, Y., Zuo, Z.: Bipartite consensus for multi-agent systems with noises over Markovian switching topologies. *Neurocomputing* **419**, 295–305 (2021)



Time Series Decision Analysis Based on Linear Programming and SARIMA

Shuai Li¹, Zeyuan Zhang²(✉), and Dongming Jiang²

¹ College of Data Science and Application, Inner Mongolia University of Technology, Hohhot, China

² College of Electric Power, Inner Mongolia University of Technology, Hohhot, China
jangomine@gmail.com

Abstract. In this paper, a decision analysis method based on time series analysis and planning model is proposed. By establishing multiple regression equations, this paper analyzes the correlation between sales volume and cost-plus pricing, and uses SARIMA model to model the time distribution characteristics of cost data and predict the future trend of data. Furthermore, this paper constructs a linear programming model to maximize the expected revenue, taking into account the constraints such as loss rate and inventory capacity. Finally, the genetic algorithm is used to solve the planning model to achieve the optimal decision of daily replenishment and pricing. The empirical analysis shows that we apply this method to the purchase strategy of supermarkets, and the results can be effectively applied to practical problems.

Keywords: Decision Optimization · Multiple Regression · Seasonal Time Series · Linear Programming · Genetic Algorithm

1 Introduction

Optimization problem is a problem that people often encounter in many fields such as real life, work and study. People are always in a certain human, material and financial conditions, looking for better or best results [1]. In general, there are many or even infinitely many schemes to be selected. The optimization method is a discipline that selects the best scheme from these feasible schemes and has achieved the best goal. According to the chronological order of events, we need to study the optimization theory based on the time series analysis [2]. For example, processing and analyzing the data sequence of gross domestic product GDP, finding the change rule of this data sequence, can well understand the development situation of China 's economy, and can accurately predict the future economic development trend, which is conducive to adjusting China 's economic measures, so as to achieve the purpose of good economic development in China [3, 4].

With the rapid development of modern computing technology and information processing technology, whether in the field of natural science, social science, military research, or engineering technology research, more and more data information needs

to be analyzed, and the technology of processing these data information becomes more and more complex with the development of the times. In these technologies, time series analysis is one of the pillars of data information technology processing. Its application in solving practical problems in various fields has received more and more attention and attention.

2 Related Works

Time series analysis can be traced back to the 1920s, even earlier in ancient Egypt 7000 years ago. The main function of time series analysis is market forecasting, and one of its main purposes is to explore the development situation of the future market. Since then, many scholars have conducted in-depth and detailed research on the theory of time series analysis and the application of solving practical problems. Since the economic system is mostly a dynamic system and is in a non-equilibrium state for a long time, dynamic optimization related to time series has emerged [5]. Dynamic programming, maximum principle, variational method and inter-temporal optimization are the basic contents and common methods of optimal control theory, and also the theoretical basis of dynamic optimization [6]. For the inter-temporal optimization problem, there are many studies at home and abroad, and fruitful results have been achieved.

As early as 1630, Galileo began to study the variational method. With the application of Euler-Lagrange theorem, Pontryagin's maximum principle, Berman equation and other principles, the theory of dynamic optimization is becoming more and more perfect [7, 8]. Nowadays, the research methods of inter-temporal optimization problems are increasingly rich [9]. There are variational method, maximum principle, dynamic programming, Hamilton-Jacobi-Bellman equation and other methods, including phase diagram analysis, comparative dynamic analysis, comparative static analysis and other methods [10]. At present, variational method and dynamic programming have been deeply studied in many disciplines such as astronomy, architecture, management, physics and so on. This theory is used to solve a series of specific problems such as resource allocation, equipment renewal, production inventory management and so on [11].

In this paper, the pricing and replenishment decisions of vegetables in fresh produce superstores were studied through multiple regression, seasonal time series, linear programming and genetic algorithms [12]. The article established multiple regression equations to analyze the relationship between total sales, cost markup and pricing of vegetable categories, analyzed the cost data by using SARIMA time series model, and finally realized the optimal decision of vegetable ordering and pricing by using genetic algorithm solution. This strategy not only considers the pricing strategies influenced by consumer behavior and market dynamics as outlined by Dong et al. and Xu and Akcay, but also optimizes the replenishment strategies to reduce wastage as demonstrated by Ketzenberg and Ferguson [13]. The goal of this study is to formulate a model that effectively balances cost, freshness, and customer demand to ensure optimal stock levels and pricing, thereby maximizing profitability and reducing environmental impact due to reduced waste.

3 Methods

3.1 Polynomial Regression Model

Polynomial regression is a machine learning technique used in regression analysis to establish the relationship between the independent variables (characteristics) and the dependent variable. The general form of polynomial regression is as follows:

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \cdots + \beta_nx_n + \varepsilon \quad (1)$$

Among them: y is the dependent variable with sales volume of each vegetable category as the dependent variable. x is the independent variable with cost plus pricing of each vegetable category as the dependent variable. $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ are the coefficient of the model, representing each polynomial coefficient. n is the polynomial order, representing the highest power of the polynomial. ε is the error term, representing the random noise that is not explained by the model.

3.2 Time Series SARIMA Algorithm

A week is a relatively short period of time that can be analyzed by analyzing the distribution pattern of recent costs over time [14]. A time series usually has many characteristics in practice, such as seasonality, trend, periodicity, stochastic fluctuations, and so on. There are also some parameters to analyze it to determine whether it meets a certain characteristic or not, and choose the appropriate model for modeling. Usually, time series can be analyzed by APIMA with three parameters (p, d, q). p for p – AR (autoregressive) order, which is the autoregressive part of the model. It allows us to compare the current value of the time series with its previous p lagged values. The autoregressive part of the model can be expressed as

$$y_t = \phi_1y_{t-1} + \phi_2y_{t-2} + \cdots + \phi_py_{t-p} + e_t \quad (2)$$

y_t is the value of the time series at time t value of the time series. q – MA (The order of the (moving average) of the

$$e_t = \theta_1e_{t-1} + \theta_2e_{t-2} + \cdots + \theta_qe_{t-q} + w_t \quad (3)$$

3.3 Linear Programming Establishes an Optimization Model

Calculations were made to find the average rate of loss for the six categories of goods, denoted as α , as shown in Table 1.

Determining decision variables: this paper defines the following decision variables in order to facilitate the model to solve for cost-plus pricing:

$$K = (1 + \gamma) \times D \quad (4)$$

Table 1. Average wastage rate

Category name	Average wastage rate α
Philodendron	12.80
Eggplant	6.68
Aquatic rhizomes	13.65
Capsicum	9.25
Edible mushroom	9.45
Cauliflower	15.52

Among them: γ is the markup rate, and D is the unit product cost, and K is the whole sale price per unit of product, and α is the wastage rate. Construct the objective function and set the maximum return as W .

$$W = (K - D) \times H \quad (5)$$

$$D = (1 + \alpha) \times D^* \quad (6)$$

Among them: W is the maximum benefit, and D^* represents the cost obtained through time series forecasting, and H represents the sales volume. Then establish the constraints: combining the background of the topic with the real life can be seen that.

- Constraint 1: Products are not considered for return cases.
- Constraint 2: Each category may be limited by stock capacity, which in general should be greater than the total daily intake. Let the maximum stock quantity be, and the total incoming quantity of each vegetable category is F .
- Constraint 3: Attrition rate $\alpha > 0$.
- Constraint 4: Total sales by category Y_{max} . Influenced by season, promotional activities and many other market factors.

Then the superstore gain is maximized as:

$$W_{max} = \{(1 + \gamma) \times D - (1 + \alpha) \times D^*\} \times H \quad (7)$$

$$f(x) = \begin{cases} 0 < \alpha \leq 1 \\ 0 \leq H \leq Y_{max} \\ y_1 = 33.74 + 0.16362039x + -0.00000835x^2 \\ y_2 = -0.87 + 0.10156872x \pm 0.00000963x^2 \\ y_3 = -3.14 + 0.12212226x + -0.00000026x^2 \\ y_4 = 0.74 + 0.11153479x + -0.00004485x^2 \\ y_5 = 29.34 + 0.11555793x + -0.00000503x^2 \\ y_6 = 24.20 + 0.11122922x + -0.00000075x^2 \\ F \leq Tmax \end{cases} \quad (8)$$

4 Results and Discussions

4.1 Data Source

The data in this article are based on the actual sales data collected and related research literature. Sales data includes the total sales of six categories of vegetables from July 1, 2020 to June 30, 2023, which have been collated and analyzed to forecast future sales trends. This paper collects and consolidates 36 months of vegetable sales data [15].

4.2 Data Optimization and Determination of Sample Size

The datasets were firstly merged, and then relevant formulas were collected to calculate the cost-plus pricing for each vegetable category. The total sales volume of six categories of vegetables for nearly 36 months from July 1, 2020 to June 30, 2023 is predicted by linear regression. The analysis yields a bar chart of the mean square error as in Fig. 1.

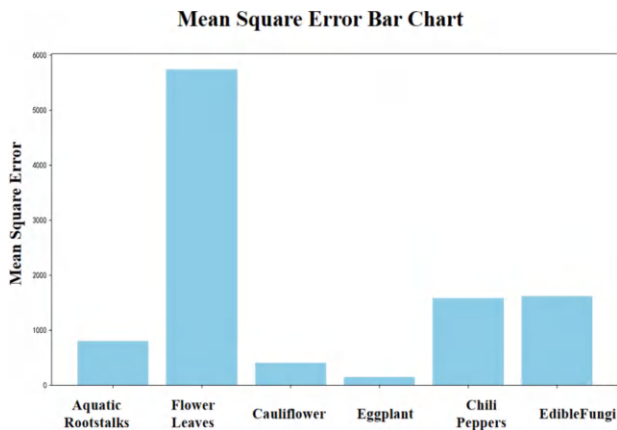


Fig. 1. Mean Square Error Bar Chart

From the figure is easy to get, flower and leaf category goods sales volume and cost-plus pricing of the worst fitting effect, so analyze the relationship between the total sales volume of each vegetable category and cost-plus pricing, do not use the linear regression model, by each vegetable category and single product sales indicators and cost-plus pricing with obvious linear relationship, this paper adopts the polynomial regression model to explore the link between the two. Based on the change rule of the total sales volume of the six vegetable categories over time in the past three years, the above linear regression model is used to predict the sales volume of July 1 ~ July 7, 2023 each category of vegetables. The predicted sales volume is shown in Table 2.

Observing the table, it can be seen that the prediction is obtained from July 1, 2023 to July 7, 2023 in the vegetable category of goods, flowers and leaves category sales are the most, followed by pepper category, eggplant sales are predicted to be the least; in addition, it is found that the predicted value of the sales volume of the same

Table 2. Forecasted sales

Date	Aquatic rhizomes	Philodendron	Cauliflower	Eggplant	Capsicum	Edible mushroom
7.1	40.058	183.487	30.420	16.152	117.182	76.935
7.2	40.062	183.485	30.405	16.150	117.240	76.946
7.3	40.067	183.483	30.391	16.141	117.298	76.957
7.4	40.072	183.480	30.376	16.132	117.356	76.968
7.5	40.077	183.478	30.362	16.122	117.414	76.979
7.6	40.082	183.476	30.347	16.113	117.472	76.990
7.7	40.087	183.473	30.332	16.104	117.530	77.001

vegetable category tends to be a constant value, which means that the total sales volume of each vegetable category is the same every day, which is obviously not in line with the reality. To analyze the reason for this, the sample data taken may be too large, and it is unreasonable and unrepresentative to analyze the sales data of the six major categories of vegetables for nearly three years in order to predict the daily sales volume on a monthly basis. In addition, due to 2021 to 2023, the country is in the new crown pneumonia epidemic period, given the closure of the city and many other factors can have an impact on all aspects of vegetable commodities indicators, which in turn affects the subsequent modeling, solving, analysis. Therefore, the following modeling, solving, testing, the sample data are taken from the February 2023 ~ June 2023 nearly five months of relevant data.

4.3 Polynomial Regression Model Results

Next, a polynomial regression model is created and fitted to extract sales data and cost-plus pricing data for each vegetable category, and the output of the fitted polynomial function corresponding to each vegetable category is shown in Table 3.

Table 3. List of Relational Curve Functions

Vegetable category	Relational curve function
Philodendron	$y = 33.74 + 0.16362039x + -0.00000835x^2$
Cauliflower	$y = -0.87 + 0.10156872x + -0.00000963x^2$
Aquatic rhizomes	$y = -3.14 + 0.12212226x + -0.00000026x^2$
Eggplant	$y = 0.74 + 0.11153479x + -0.00004485x^2$
Capsicum	$y = 29.34 + 0.11555793x + -0.00000503x^2$
Edible mushroom	$y = 24.20 + 0.11122922x + -0.00000075x^2$

It is proposed to merge and plot the relationship between total sales and cost-plus pricing for each vegetable category. The results are shown in Fig. 2.

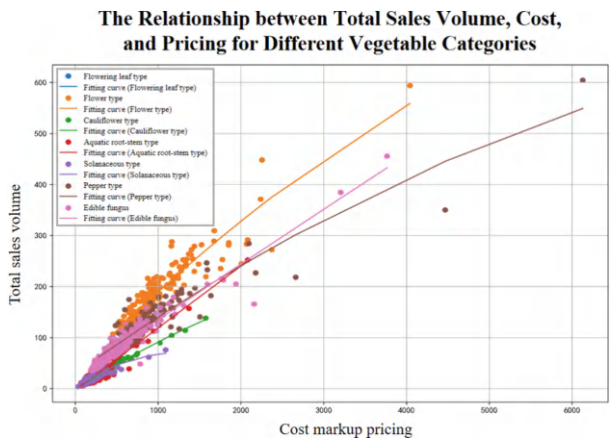


Fig. 2. Relationship between total sales and pricing of different vegetable categories

Coefficient of determination R^2 : The value is used to indicate how much of the variability in the dependent variable is explained by the model and is a measure of the quality of the regression model, which indicates the percentage of the variance of the data that is explained by the model. R^2 ranges from 0 to 1, where 1 indicates that the model explains the data perfectly and has a good fit. The coefficient of determination was calculated for each vegetable category as shown in Table 4.

Table 4. Coefficient of determination for each vegetable category

Vegetable Category	Coefficient of Determination
Philodendron	0.8554
Cauliflower	0.9460
Aquatic rhizomes	0.8934
Eggplant	0.7764
Capsicum	0.8733
Edible mushroom	0.9221

The cauliflower species possessed the highest R^2 value of 0.9460, indicating a very good fit. Most of these models showed a good fit, especially for the “Cauliflower” and “Edible Mushroom” categories. This indicates that the quadratic polynomial regression used is valid for these data. The positive and negative coefficients of the model indicate the relationship between cost-plus pricing and the total volume of sales. For example, a positive primary coefficient indicates that when prices rise, sales volume also rises,

while a negative quadratic coefficient may imply that there is an optimal price point beyond which sales volume may fall.

4.4 Time Series SARIMA Algorithm

The seasonal decomposition of each category of this data is shown in Fig. 3.

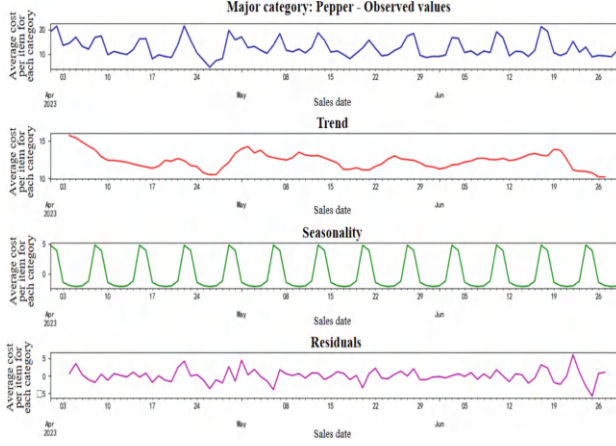


Fig. 3. Seasonal Breakdown of Chili Peppers

Observing the fourth chart, it can be seen that for most of the time, the residuals fluctuate around the zero line, which is a good sign indicating that the model does not have a systematic bias. Observation of the third seasonality plot reveals that the chili category has strong seasonality on a weekly basis, and in predicting the cost problem over a short period of time, it is not easy to observe strong regularity using three years of data. Therefore, SARIMA was used for the analysis. This is a seasonal version of the ARIMA model. The s is added to the ARIMA (p, q, d) model as a specialized model for seasonal analysis. For data with seasonality, a differencing process is performed. d is the number of times the data are differenced, and the differencing operation for a series with periodicity s is as follows:

$$\begin{aligned} W_t &= \nabla \nabla_{12} X_t \\ \nabla_s X_t &= (1 - B^s) X_t \\ \nabla_s^d &= (1 - B^s)^d X_t \end{aligned} \quad (9)$$

The model was ordered using BIC. Due to the seasonal nature of the data, the BIC criterion for ordering the heat map required repeated fitting, so the auto arima function in python's pmdarima was used to automatically select the optimal parameters. In the case of chili peppers, for example, the $p = 2, q = 1, d = 0, s =$. The model was built for prediction, and the results are shown in Table 5.

$$(1 - B)(1 - B)(1 - B^7)X_t = (1 + \theta B)\varepsilon_t \quad (10)$$

Table 5. Projected results

	Aquatic rhizomes	Philodendron	Cauliflower	Eggplant	Capsicum	Edible mushroom
1	18.08	8.06	51.14	24.81	16.43	7.92
2	19.08	7.30	44.82	24.83	13.55	6.58
3	12.50	4.71	24.72	15.03	8.90	4.17
4	13.45	4.91	25.94	12.24	9.05	4.07
5	15.69	4.97	28.10	12.88	9.17	4.29
6	18.83	4.16	27.20	13.63	11.04	4.55
7	17.10	5.00	30.16	16.85	10.13	4.87

4.5 Genetic Algorithm to Solve Daily Replenishment and Pricing Strategy

Genetic algorithm is a search and optimization algorithm based on natural selection and genetic mechanisms. It simulates the process of biological evolution in nature in order to solve optimization problems with specific objectives in artificial systems. The core idea of genetic algorithm is to improve the quality of the solution generation by generation through population search based on the principle of survival of the fittest. Genetic algorithms simulate the process of biological evolution by generating, evaluating and improving a set of solutions to find the optimal or near-optimal solution to a problem. The following is the basic solution flow of the algorithm:

- 1) Population initialization: for the initial population, the chromosomes for each individual are usually randomly generated. If a chromosome is a binary string of length n , then:

$$c_i = \{b_1, b_2, \dots, b_n\} \tag{11}$$

Among them: c_i is the first i chromosome, and b_i is a random binary bit (0 or 1).

- 2) Define the degree function: compute the fitness value for each individual. Assume that f is the fitness function

$$F_{(C_i)} = f(c_i) \tag{12}$$

Among them: $F_{(C_i)}$ is the chromosome of the fitness value of c_i .

- 3) Selection: the use of fitness values to select individuals in a population. Roulette assignment selection is a commonly used selection method with probabilities given by the following equation:

$$P_{(C_i)} = \frac{F(c_i)}{\sum_{j=1}^N F_{(C_i)}} \tag{13}$$

where N is the population size.

- 4) Crossover pairing: if chromosomes are binary coded, then a single point crossover can be described as follows, with a randomly chosen crossover point k . For two parents C_a and C_b , the child chromosome is:

$$C_{child} = \{b_1^a, b_2^a, \dots, b_k^a, \dots, b_{k+1}^b, \dots, b_n^b\} \quad (14)$$

$$C_{child} = \{b_1^b, b_2^b, b_k^a, \dots, b_k^a, b_{k+1}^b, \dots, b_n^b\} \quad (15)$$

- 5) Variation: for binary coding, variation involves randomly flipping a bit. given the mutation rate μ the probability that each bit will be flipped is: $P_{mutation}(b_j) = \mu$. For a chromosome c_i for each bit of the chromosome b_j , apply that rate above for possible flips.
- 6) Replacement: new individuals can replace old populations or combine with old populations to form new populations. A common method is “steady-state” substitution, in which only a small fraction of the worst old individuals are replaced by new ones.
- 7) Termination condition: the stopping condition can be reaching the maximum number of iterations T or the adaptation exceeds a threshold value $F_{threshold}$ or the adaptation of several consecutive generations is less than some small value, etc.

The principle of parameter design in it:

- a) Population size: the choice of population size should balance search performance and computational resources. In general, the population size is usually between 50 and 200.
- b) Crossover rate: higher crossover rates help to maintain population diversity, but too high crossover rates may lead to slow convergence. Typically, crossover rates range from 0.6 to 0.9.
- c) Mutation rate: a low mutation rate may lead to falling into a local optimum solution, while a higher mutation rate may destroy useful genes. Typically, the usual range of values is 0.5%–1%.
- d) Evolutionary generation: should not be too large or too small, generally selected 100 ~ 500.

In this paper, we have implemented Python to solve the model and determine the pricing strategy, which is brought into the regression model to find out the sales volume of each vegetable category and then determine the daily replenishment, as shown in Tables 6 and 7.

4.6 Model Comparison

Other scholars may explore more complex machine learning algorithms or optimization algorithms, such as neural networks, support vector machines, ant colony algorithms, etc. for vegetable pricing or replenishment decisions in order to improve the prediction accuracy or to seek for more efficient optimal decisions. They may also investigate how to process data in real time and build decision support systems to enable more timely decisions. In addition, they may consider risk management and robustness analysis to cope with the risks associated with market volatility.

Table 6. Pricing strategy

	Aquatic rhizomes	Philodendron	Cauliflower	Eggplant	Capsicum	Edible mushroom
1	18.08	8.06	51.14	24.81	16.43	7.92
2	19.08	7.30	44.82	24.83	13.55	6.58
3	12.50	4.71	24.72	15.03	8.90	4.17
4	13.45	4.91	25.94	12.24	9.05	4.07
5	15.69	4.97	28.10	12.88	9.17	4.29
6	18.83	4.16	27.20	13.63	11.04	4.55
7	17.10	5.00	30.16	16.85	10.13	4.87

Table 7. Daily replenishment

	Aquatic rhizomes	Philodendron	Cauliflower	Eggplant	Capsicum	Edible mushroom
1	22.22	192.43	29.61	29.98	128.49	63.92
2	21.43	188.15	26.48	32.33	102.62	46.94
3	12.96	120.82	14.32	18.67	68.33	28.39
4	15.41	134.71	14.56	14.57	69.05	30.05
5	17.43	126.74	16.54	16.23	75.95	41.15
6	23.82	107.25	16.13	15.44	85.08	40.74
7	19.37	130.84	18.06	19.04	81.52	35.83

5 Conclusions

In this study, a decision analysis framework is proposed by combining time series analysis and planning model. The SARIMA algorithm effectively captures the temporal dynamics and seasonal characteristics in the data, providing a basis for trend prediction. Then, a linear programming model is established to reflect the decision-making process, add constraints, and maximize the expected results. Finally, the near-optimal solution is identified by genetic algorithm.

However, there are still limitations in this study. The accuracy of the model is very dependent on historical data, and may not fully consider the impact of sudden environmental changes on decision-making. In addition, the application of the model in different products and markets has not been widely tested.

Future research needs to explore the robustness of the model in different situations, integrate real-time data streams to provide more dynamic decision support, and consider other optimization algorithms to help the model improve the quality and computational efficiency of the solution.

Author Contributions. Shuai Li, Zeyuan Zhang and Dongming Jiang are co-first authors.

References

1. Cai, X.Q., Chen, J., Xiao, Y.B., Xu, X.L.: Optimization and coordination of fresh product supply chains with freshness keeping efforts. *Prod. Oper. Manag.*, To appear (2009)
2. Ketzenberg, M., Ferguson, M.: Managing Slow Moving Perishables in the Grocery Industry. *Prod. Oper. Manag.* **17**(5), 513–521 (2008)
3. Cai, L.Q., Chen, J., Yan, H.M.: Single-period two-product inventory model with substitution: solution and properties. *J. Syst. Sci. Syst. Eng.* **13**(1), 190–201 (2004)
4. Yao, D.D., Zheng, S.: Inventory with substitution: single and multi-period models. In: Shan-thikumar, J.G., Yao, D.D., Zijm, W.H.M. (eds.) *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, Kluwer, Chapter 8, pp. 177–202 (2003)
5. Bassok, Y., Anupindi, R., Akella, R.: Single-period multi-product inventory models with substitution. *Operations Research* **47**(4), 632–642 (1999)
6. Dong, L., Narasimhan, C., Zhu, K.: Product line pricing in a supply chain. *Manage. Sci.* **55**(10), 1704–1717 (2009)
7. Talluri, K.T., Van Ryzin, G.J.: Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1), 15–33 (2004)
8. Ben-Akiva, M., Lerman, S.: *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press (1985)
9. Tirole, J.: *The Theory of Industrial Organization*. MIT Press (1988)
10. Xu, S., Akcay, Y.: Joint dynamic pricing of multiple perishable products under consumer choice. working paper, Smeal College of Business. The Pennsylvania State University (2008)
11. Federgruen, A., Heching, A.: Combined pricing and inventory control under uncertainty. *Oper. Res.* **47**(3), 454–475 (1999)
12. Zhang, Y.: Optimization of live e-commerce supply chain ordering strategy for seasonal demand. *Henan Univ. Technol.* (2023). <https://doi.org/10.27791/d.cnki.ghegy.2023.000763>
13. Luo, Z., Hong, Y.: Research on production, ordering, and pricing of single substitution product supply chain under capacity constraints. *Operat. Res. Manage.* **31**(07), 186–192 (2022)
14. Hu, Y.: Research on coordination of fresh supply chain considering preservation efforts and value-added services under retailer risk aversion. Fuyang Normal University (2022). <https://doi.org/10.27846/d.cnki.gfysf.2022.000121>
15. Shi, P.: Research on inventory management of agricultural products at different stages. Inner Mongolia University of Finance and Economics (2022). <https://doi.org/10.27797/d.cnki.gnmgc.2022.000199>



Artificial Intelligence-Driven Network Intrusion Detection and Response System

Haokun Chen¹, Yiqun Wang¹(✉), Shangyu Zhai¹, Wanrong Bai², Zhiqiang Diao³,
and Dongyang An¹

¹ Gansu University of Political Science and Law, Lanzhou, Gansu, China
wyq6696@gsupl.edu.cn

² State Grid Gansu Electric Power Research Institute, Lanzhou, Gansu, China

³ Tianjin Public Security Bureau, Tianjin, China

Abstract. Data moving across networks has increased thanks to technological developments, which has also created network security issues. The sophistication of these threats makes conventional intrusion detection systems unable to adequately manage these fresh security issues. Consequently, since neural-based intrusion detection systems can manage dynamic and complicated data, focus has turned to leveraging them to address these challenging cyber security issues. This work investigates how neural-based intrusion detection systems address challenging cyber security issues. Thus, for this work, important neural-based models of relevance are CNN and MLP models that have been shown to be useful in capturing abnormalities in big datasets and so pertinent for intrusion detection systems. The NSL-KDD dataset was applied for the experiment and served to support the idea on the application of neural-based models in intrusion detection. Deep learning models show, based on their accuracy, F1-score, and recall values, far higher performance in spotting anomalies than conventional models. These findings allow one to deduce that neural-based intrusion detection systems present a strong method for network the detection and prevention of cyber security events.

Keywords: Intrusion Detection System · Deep Learning · Neural Networks · CNN · MLPs

1 Introduction

Technological evolution has seen an increase in devices sharing data with networks and this has consequently led to an increase in malicious attacks targeting these networks. These constantly evolving attacks pose a threat to organizations as traditional intrusion detection systems (IDS) often struggle to handle these new evolving threats [1]. Because of these dynamic threats, this paper presents a novel artificial intelligence (AI)-driven network intrusion detection and response system, leveraging cutting-edge AI algorithms, including deep learning and machine learning, to enhance the efficacy of network security measures.

The use of machine learning to improve intrusion detection is not a new phenomenon as studies have previously looked into the subject. For instance, Gadai et al. propose a

hybrid approach that uses k-means clustering to improve the performance of intrusion detection systems, especially when dealing with high-dimensional and complex network traffic patterns [2]. With the rapid innovation and development of network technologies, there is an immediate need for an IDS capable of detecting threats on a real-time basis.

Major Contributions:

- (1) Introducing a robust AI-driven system that transcends the limitations of traditional IDS by employing advanced algorithms capable of real-time monitoring and adaptive threat identification.
- (2) Demonstrating superior detection accuracy, recall, and F1 value through the deployment of a deep learning-based IDS, outperforming conventional methods.
- (3) Engineering a system that not only detects but also actively responds to intrusions, thereby fortifying the proactive defense of broadband networks.
- (4) Conducting an in-depth review of the current cyber threat landscape, highlighting the intelligent and insidious nature of modern attacks and underscoring the urgency for advanced defense strategies.

The remainder of this article is structured as follows: Sect. 2 is for literature review and this is where a review of relevant literature occurs, especially on deep learning model usage in IDS. Section 3 will cover the methodology section, the section where the experimental set-up is discussed including the dataset, model configuration and evaluation metrics. Section 4 covers the discussions and findings from the experiment. Here, the model performance for the proposed deep learning model is compared with that of traditional models. Section 5 is conclusion and here, the findings of the study are summarized, limitation of the research stated and areas for future research proposed.

2 Literature Review

Neural networks are increasingly becoming a powerful tool in intrusion detection systems because they are more effective in detecting intrusions. The effectiveness of neural based IDS systems is a result of their ability to account for nonlinear relationships between variables and also to learn from large datasets. These key characteristics have neural networks more powerful unlike traditional rule-based IDS systems in network environments where has been increasing exponentially and also threats constantly evolving.

2.1 Neural Networks in Intrusion Detection Systems

The rise of network-based applications like Internet of Things (IoT) have led to an increase in data flowing through networks and consequently cyber threats targeting these networks. In their article on deep learning based IDS, Lansky et al. [3] & Naskath et al. [4] states that deep learning models like Convolutional Neural Networks (CNNs) and Multilayer Perceptron's (MLPs) among other models have proven effective in addressing the challenges that have been associated with rule-based systems that traditional IDS systems used. The success of these models in addressing the limitations of rule-based models makes them a good choice for IDS systems.

Maithem & Al-sultany [5] in their article discusses the use of deep neural networks in detecting anomalies concerning data breaches and they propose a deep learning model to help detect abnormal traffic in network traffic. Below is an illustration of their proposed deep neural network for IDS.

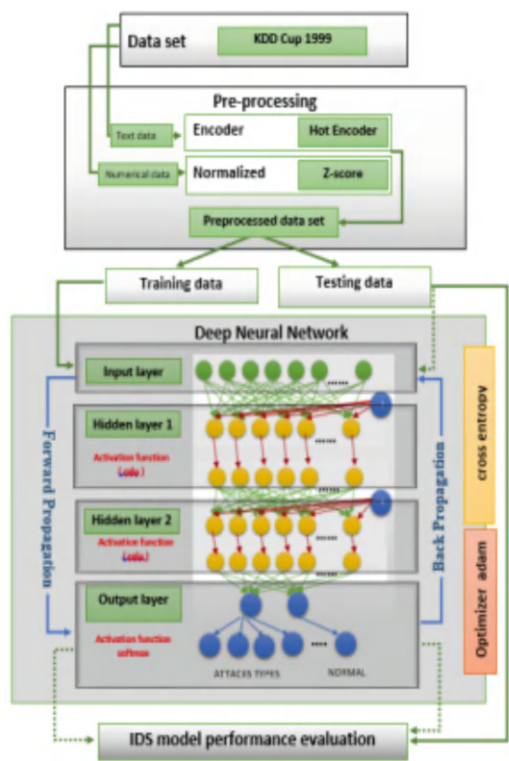


Fig. 1. Proposed DNN-IDS

In Fig. 1 above, Maithem & Al-sultany [5] conceptualized a deep learning model that takes network traffic data as input and outputs classification whether the traffic is normal or not.

2.2 Nonlinear Relationships in Network Data

A key limitation of rule-based models in IDS is that they are incapable of capturing nonlinear relationships in data. Chen et al. [6] posits that cyberattacks manifest nonlinear behavioral patterns, which neural networks are capable of capturing. The weighted connections between layers in neural networks facilitate the automatic learning and revelation of data’s underlying nonlinear characteristics. This feature learning ability allows for the detection of attack patterns that simple linear models cannot discern, such as the progressive attack behaviors that may be hidden within typical traffic flows. By effectively managing the complex relationships concealed in data, neural networks improve

the detection capabilities for unknown and latent attacks and adapt to various intrusion patterns.

2.3 Dynamic Adaptations and Real-Time Learning

Neural based models that are best placed to handle constantly evolving threats as they are capable of learning dynamically and detect the threats in real time. Hname & Hussain [7] argue that neural networks have the agility to update their rules in lockstep with the live traffic flow of the network. For instance, with the emergence of novel attack vectors, the system can swiftly recalibrate its rules to ensure that intrusion detection mechanisms are primed to spot threats at the earliest opportunity. A feedback loop embedded within the neural network architecture ensures ongoing optimization of detection rules, thereby enhancing the system's responsiveness and the precision of its detection capabilities.

2.4 Time-Series Analysis in Threat Prediction

Cyberattacks are usually interconnected series of events and not singular activities. Neural networks grounded in time series analysis can use the interconnectedness of events to anticipate attack patterns and detect potential threats in advance. In their article on neural networks in time series, Zhang et al. [8] posit that network models have found usage in stock prediction because they are efficient in processing non-linear time series data. Neural network models in time series helps in identifying correlations in system behaviors. Specifically, neural network models like Long Short-Term Memory networks (LSTMs) and Recurrent Neural networks (RNNs) have been found to be exceptional in capturing long-term dependencies in actions within a system and this is key in identifying attacks that are on an incremental nature [9].

2.5 Performance of Neural Network Models

A good performing model is critical for the success of intrusion detection systems. Model performance can change over time but to ensure that the model performs excellently, it is important to optimize the model. Continuous model optimization through adjusting model parameters to reduce cases of overfitting and improve the error detection rates of the model [10]. Sajid et al. (2024) argues that to mitigate overfitting, it is important to use regularization techniques as they will ensure that the model has good generalization on new data. Because of such approaches, the final model will effectively ingest voluminous and complex data and make accurate predictions.

From the reviewed literature, it is evident that scholars view neural-based IDS as the future of models for use in intrusion detection systems because they are capable of handling nonlinear relationships and also their dynamic nature. These of these models guarantees better performance and accuracy in detecting anomalies in network data. Therefore, as threats continue evolving, adopting neural-based IDS models and optimizing them is essential in having well-performing models.

3 Methodology

The approach to be applied in creating and assessing the neural models suggested for usage in intrusion detection systems is contained in this part. To verify the effectiveness of the model, the experiments was designed and evaluated using the NSL-KDD data. From the analysis, we will come up with conclusions on the proposal on the use of neural-based models in intrusion detection systems. We also provide a detailed summary of the dataset, the experiment set up, results and a comprehensive evaluation.

3.1 Selection and Description of Dataset

In experiments involving deep or machine learning models, it is always important to choose a dataset that has features related to the problem at hand and also of good quality. In this case, the NSL-KDD dataset was chosen because of the good quality and the also previous usage of the dataset in intrusion detection problems. Below is the specific description of the dataset:

Dataset Name: NSL-KDD (KDD Cup 99) Dataset*

3.1.1 NSL-KDD Dataset

The dataset is has been widely used in intrusion detection problems and that is why it was chosen for this experiment. Even though the data has issues like imbalanced classes or redundant samples, these issues can be addressed through the removal of the redundancies and optimizing the class distribution to make it suitable for use in machine and deep learning models.

3.1.2 Dataset Functionality

The dataset has diverse network attacks and this will help the model learn and differentiate between normal and abnormal network traffic. Additionally, each of the observations in the dataset has label indicating whether the activity is normal or not.

Feature dimension: Contains 41 features, including the starting and ending points of the connection, protocol type, service type, flag, etc.

Data volume: The dataset size is moderate and suitable for training deep learning models, ensuring that the models have good generalization ability.

3.1.3 Application Scenarios of the Dataset

Anomaly detection: Using supervised learning models to explore abnormal behavior in large-scale network data.

Classification problem: Classify multiple types of network attacks and evaluate the model's ability to recognize different types of attacks.

The application of deep learning: High dimensional features are suitable for studying deep learning models such as neural networks and convolutional neural networks (CNN) to improve detection performance.

The general process of intrusion detection includes information collection, data pre-processing, detection and analysis, and responding according to security policies. The

NSL-KDD dataset covers various data types required for these steps, providing a solid foundation for this study.

3.2 Experimental Setup

To ensure the effectiveness and reproducibility of the experiment, this study has made detailed configurations in the following aspects:

3.2.1 Parameter Configuration

Reasonable configuration of model parameters and training parameters is the key to ensuring the reliability of the results during the experimental process. The specific parameter configuration is as follows:

Deep learning model parameters:

Network architecture: Two different architectures, Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN), are used to compare their performance in intrusion detection.

Layers and Number of Neurons: The MLP model consists of 3 layers, with each layer containing 64, 128, and 64 neurons; The CNN model consists of two convolutional layers, each containing 32 and 64 filters, followed by a fully connected layer.

Activation function: Use ReLU activation function to improve the non-linear expression ability of the model.

Learning rate: set to 0.001, adjusted through validation set to optimize model convergence speed and effectiveness.

Training parameters:

Dataset partitioning: Divide the NSL-KDD dataset into training, validation, and testing sets in a ratio of 70:15:15 to ensure consistent performance of the model on different datasets.

Below is a summary of the model training parameters:

Table 1. Model parameters

Parameter	MLP	CNN
Network Architecture	Multi-Layer Perceptron	Convolutional Neural Network
Layers and Neurons	3 layers (64, 128, 64)	2 convolutional layers (32, 64), 1 fully connected layer
Activation Function	ReLU	ReLU
Learning Rate	0.001	0.001
Training Epochs	50	50
Batch Size	32	32
Optimization Algorithm	Adam	Adam
Dropout Rate	0.5	0.5

Table 1 above summarizes the parameters to the models used for this project.

In addition to the neural models mentioned, we found it fair to compare the performance of these models with a hybrid data mining method that Gadai et al., proposed. This model combined k-means clustering with sequence minimum optimization (SMO) classification algorithm which was used as a benchmark for comparative analysis [2].

3.2.2 Experimental Environment

Below are the software and hardware environment used for this experiment:

Hardware configuration:

Processor: Intel Core i7-10700 K CPU

Memory: 32 GB DDR4

GPU: NVIDIA GeForce RTX 3080

Storage: 1 TB SSD for fast reading and storage of datasets

Software environment:

Operating System: Ubuntu 20.04 LTS

Deep learning frameworks: PyTorch 1.10 and Tensor Flow 2.8 and PyTorch 1.10

Programming language: Python 3.8, Scikit learn

Other tools: Jupyter Notebook

3.3 Experimental Results and Analysis

In this section, we present the experimental results of the neural-based IDS and also a comparison with the benchmark model.

3.3.1 Intrusion Detection Performance Evaluation

To comprehensively evaluate the model performance, we used various performance indicators accuracy, recall, F1 score, and precision. Below is a table showing the model performance (Table 2).

Table 2. Model performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
K-means + SMO)	85.2	82.5	80.3	81.4
MLP	89.7	87.2	85.5	86.3
CNN	92.4	90.1	88.7	89.4

Result analysis:

With regard to accuracy, the CNN model excelled at 92.4%, much above the benchmark performance of 85.2%.

The good result indicates that CNN model can identify trends in complicated data.

CNN has clearly shown quite performance on accuracy and recall at 90.1% and 88.7% respectively. CNN performs rather well in precisely spotting normal and aberrant behavior.

In the evaluation of classification models, F1-score is a crucial indicator since it indicates how well the model strikes between recall and accuracy. Here is F1 score's formula.

F1-score is an important metric in classification model evaluation as it shows how the model balances between recall and precision. Below is the formula for F1 score.

$$F1 - score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (1)$$

On F1-score, CNN scored highly at 89.4% which is still significantly higher compared to the benchmark score of 81.4%. This shows that CNN model is excellent in balancing both accuracy and recall.

Comparison with existing methods:

By comparing with Gadal et al.'s hybrid data mining method, it can be seen that the neural network-based approach significantly improves various indicators [2]. This is mainly due to the advantages of neural networks in feature learning and nonlinear relationship modeling, which enable them to more effectively capture complex intrusion patterns.

3.3.2 Effectiveness Evaluation of Response System

In addition to performance testing, this study also evaluated the effectiveness of the response system, including response time, real-time performance, system stability, and scalability.

Response time:

The average response time of the benchmark method is 200 ms.

The response time of the MLP model is 150 ms, while the CNN model further reduces it to 120 ms.

Real time performance:

The CNN model is capable of processing over 1000 packets per second in real-time environments, meeting the requirements of high traffic networks.

System stability:

Long term running tests have shown that the neural network model has not experienced memory leaks or performance degradation within 72 h of continuous operation, demonstrating good stability.

Scalability:

When dealing with large-scale datasets, CNN models exhibit good scalability and can linearly scale by increasing computational resources as the data volume increases, with minimal performance loss.

Result Discussion:

Response time: Neural network models, especially CNN, significantly shorten response time and improve the system's real-time detection and response capabilities. This is particularly important for network environments that require rapid response.

Real time performance: High processing speed ensures the feasibility of the system's application in high traffic networks, avoiding security vulnerabilities caused by latency.

System stability: Long term stable operation indicates that the neural network model has good robustness and can work reliably in practical applications for a long time.

Scalability: Good scalability enables neural network models to adapt to the constantly growing network size and data volume, ensuring that the system remains competitive in future development.

4 Detailed Discussion and Analysis

Based on the above experimental results, the following conclusions can be drawn:

4.1 The Superiority of Neural Networks

Neural networks, especially CNNs, perform better in intrusion detection than traditional hybrid data mining methods. Its powerful feature learning and nonlinear modeling capabilities enable it to more accurately identify complex intrusion patterns.

4.2 The Impact of Model Selection

CNN captures local features and handles high-dimensional data better than MLP, hence enhancing detection performance. This implies that in intrusion detection activities, choosing an appropriate neural network design is quite essential.

4.3 The Effect of Dynamic Rule Adjustment

Although the experiment did not specifically assess the independent effect of dynamic rule adjustment, it can be deduced from the general performance improvement that the mechanism has helped to adapt the model to new attacks and modify network environments.

4.4 Feasibility of Practical Application

The great accuracy and low response times suggest that practical network settings could find use for intrusion detection systems built on neural networks. Its scalability and stability help to improve its practical relevance even further.

4.5 Future Optimization Direction

Although in this study neural networks have shown good performance, opportunity for improvement exists. For instance, adding more sophisticated technologies (like attention processes) to improve the detection capacity of the model or further optimising the network structure to lower computational resource consumption.

This chapter uses thorough experimental design and evaluation to validate, via network security, the efficiency of neural network-based intrusion detection techniques. With regard to accuracy, response speed, and system stability, the experimental results reveal that neural networks—especially CNN models—outperform conventional techniques, so attesting to their great advantages. Concurrent with this, the exceptional handling of complicated network incursion patterns by neural networks has been verified by means of comparison study with current approaches. This lays a strong basis for next studies and useful applications and shows the path of additional optimisation and enhancement.

5 Conclusion

This work presents a set of network intrusion detection and response system based on artificial intelligence technology and accomplishes a number of significant results by means of sophisticated technologies like deep learning. The introduction of adversarial attack testing and the use of deep learning in intrusion detection reflects mostly the original inventiveness of this work. Adversarial attack results reveal that the system is rather resistant; deep learning application increases intrusion detection intelligence and adaptation capacity. On several performance standards, including accuracy, precision, and recall, the system performs satisfactorily. It has a high success rate in the adversarial attack situations and great detection capacity to several network threats. Future technical directions call for multimodal fusion, adaptive learning, and so on. These developments will assist to raise the system's intelligence, flexibility, and completeness. Apart from the subject of network intrusion detection, the application field of the system can be enlarged to include cross-domain collaboration, edge computing security, and other domains. This will increase the system's adaptability and fit for several conditions and surroundings. Furthermore lacking in this work are various constraints including reliance on specific parameter combinations and limits of the dataset. Future research with more datasets and more thorough parameter adjustment could handle these problems.

Finally, our effort has produced outstanding results in the field of network security, which offers great support for further progress of network intrusion detection and response systems. By means of ongoing technological innovation and application extension, the performance and adaptability of the system can be enhanced even further to better address the development obstacles of network security.

References

1. Asgharzadeh, H., Ghaffari, A., Masdari, M., Gharehchopogh, F.S.: An intrusion detection system on the internet of things using deep learning and multi-objective enhanced gorilla troops optimizer. *J. Bionic Eng.* **21**(5), 2658–2684 (2024). <https://doi.org/10.1007/s42235-024-00575-7>
2. Gadal, S., Mokhtar, R., Abdelhaq, M., Alsaqour, R., Ali, E.S., Saeed, R.: Machine learning-based anomaly detection using K-mean array and sequential minimal optimization. *Electronics* **11**(14), 2158 (2022). <https://doi.org/10.3390/electronics11142158>
3. Lansky, J., et al.: Deep learning-based intrusion detection systems: a systematic review. *IEEE Access* **9**, 101574–101599 (2021). <https://doi.org/10.1109/access.2021.3097247>
4. Naskath, J., Sivakamasundari, G., Begum, A.A.S.: A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN. *Wireless Pers. Commun.* (2022). <https://doi.org/10.1007/s11277-022-10079-4>
5. Maithem, M., Al-sultany, G.A.: Network intrusion detection system using deep neural networks. *J. Phys: Conf. Ser.* **1804**(1), 012138 (2021). <https://doi.org/10.1088/1742-6596/1804/1/012138>
6. Chen, S., Wu, Z., Christofides, P.D.: Cyber-attack detection and resilient operation of nonlinear processes under economic model predictive control. *Comput. Chem. Eng.* **136**, 106806 (2020). <https://doi.org/10.1016/j.compchemeng.2020.106806>
7. Hnamte, V., Hussain, J.: Dependable intrusion detection system using deep convolutional neural network: A Novel framework and performance evaluation approach. *Telematics and Informatics Reports* **11**, 100077 (2023). <https://doi.org/10.1016/j.teler.2023.100077>

8. Zhang, L., et al.: Time-series neural network: a high-accuracy time-series forecasting method based on kernel filter and time attention. *Information* **14**(9), 500 (2023). <https://doi.org/10.3390/info14090500>
9. Psychogyios, A.P., Bourou, S., Nikolaou, N., Maniatis, A., Zahariadis, T.: Deep Learning for Intrusion Detection Systems (IDSs) in Time Series Data. *Future Internet* **16**(3), 73 (2024). <https://doi.org/10.3390/fi16030073>
10. Sajid, M., et al.: Enhancing intrusion detection: a hybrid machine and deep learning approach. *J. Cloud Comp. Adv. Sys. Appl.* **13**(1) (2024). <https://doi.org/10.1186/s13677-024-00685-x>



Identifying Consumer Behavior Patterns from Massive User Transaction Data Based on Data Mining Techniques

Qi Wang^(✉)

School of Economics and Management, Dongying Vocational College of Science and Technology, Dongying 257300, Shandong, China
chinwangqi@126.com

Abstract. In response to the urgent need for consumer behavior analysis and the complexity of massive transaction data, this article conducts in-depth research on user transaction data based on data mining techniques to identify their potential consumption behavior patterns. Firstly, the article conducts data preprocessing operations on the massive transaction data collected; subsequently, key attribute features of the user are generated through feature extraction. Next, the article uses the K-means clustering algorithm to cluster users' consumption behavior, in order to divide user groups into different consumption types. On this basis, the Apriori algorithm is applied to mine the association rules of user groups and identify the consumption preferences and linkage behaviors that exist between specific groups. At the same time, in order to capture the dynamic characteristics of consumer behavior, the article also uses the ARIMA (Autoregressive Integrated Moving Average) time series model to analyze the cyclical and seasonal trends of consumer behavior. Finally, by constructing a classification model based on random forests, the article predicts future consumer behavior. The experimental results show that in the user clustering experiment, the clustering effect is best when $K = 4$, with an average contour coefficient of 0.73; in association rule mining experiments, higher support and confidence can effectively improve the accuracy of rules; in the time series prediction experiment, the Mean Squared Error (MSE) of the ARIMA model is 0.8421, and the Mean Absolute Percentage Error (MAPE) is 7.63%. The comprehensive data mining method proposed in this article can effectively identify and predict users' consumption behavior patterns, significantly improving the accuracy of consumption behavior analysis and proving the advantages and feasibility of the adopted method in analyzing massive transaction data.

Keywords: Data Mining · Cluster Analysis · User Transaction Data · Association Rules · K-Means Clustering Algorithm

1 Introduction

With the rapid development of e-commerce and the popularity of online transactions, the scale of consumer behavior data generated by users is becoming increasingly large. These massive transaction data contain rich consumer behavior information. Digging

deeper into these data can not only help businesses better understand users' consumption habits, but also provide important basis for personalized recommendations and precision marketing. However, traditional data analysis methods often struggle to effectively identify hidden behavioral patterns when faced with such massive and complex consumer data. Therefore, utilizing advanced data mining techniques to extract valuable consumer behavior patterns from massive user transaction data has become an important research topic in business decision-making and market analysis.

This article proposes a comprehensive method based on data mining technology to analyze users' consumption behavior patterns. By preprocessing user transaction data, extracting features, and using K-means clustering algorithm to identify different consumer groups, the Apriori algorithm is further applied to mine users' consumption association rules. In addition, this article also combines the ARIMA time series model to analyze the dynamic changes in consumer behavior, and finally predicts future consumer behavior through a random forest model. This method effectively integrates clustering, association rule mining, and time series analysis, and can comprehensively capture users' consumption behavior characteristics, providing data support for subsequent personalized recommendations and business strategies.

The structure of this article is as follows: The first part introduces the background and research significance of consumer behavior analysis, and summarizes existing related research results; the second part provides a detailed description of the data preprocessing methods, feature extraction process, and various algorithms used; the third part presents the experimental design and results, and analyzes and discusses the effectiveness of the method; the final section summarizes the main contributions of the research and proposes possible future research directions.

2 Related Works

Currently, many scholars are attempting to use data mining techniques to analyze user consumption behavior. For example, Anita and Patil [1] aim to apply business intelligence to the retail industry by providing relevant and timely data to identify potential customers. These data are based on systematic research and scientific applications, used to analyze consumers' sales history and purchasing behavior. Aziz and Aftab [2] aimed to use data mining techniques to rank the three diets of respondents in the dataset, and explored the advantages and limitations of these tools, such as the need for extensive manipulation before analysis and the ability to obtain consistent results even after eliminating analysis bias. Ageed et al. [3] discussed the fusion applications of cloud computing, data mining, and large-scale online data. They studied methods for big data mining in cloud systems and proposed solutions to the issues of cloud compatibility and computing technology to promote the development of big data mining in cloud systems. Moinuddin et al. [4] delved into the role of marketing analysis in interpreting consumer behavior and driving event success, and studied the complex interrelationships between marketing analysis techniques, consumer insights, and event performance indicators through a mixed methods approach. Safara [5] proposed a model that uses machine learning methods to predict consumer behavior and tested the performance of five independent classifiers and their integration with Bagging and Boosting on a dataset collected from online shopping

websites. Data mining is typically defined as the technique of discovering meaningful patterns or interesting information for decision support in large amounts of data. Santoso [6] hoped to improve subsequent sales strategies by using this prior algorithm to search for patterns. Liao et al. [7] studied Taiwanese users and used data mining methods to generate user profiles through cluster analysis. They combined association rule analysis to investigate the development of social media and social applications in social network interaction, proposing two patterns and several important rules. Saura [8] summarized recommendations for developing digital marketing strategies for businesses, marketers, and non-technical researchers, and outlined future research directions for innovative data mining and knowledge discovery applications. Overall, existing research still has significant shortcomings in addressing data dimensionality and noise processing.

Some studies have shown that data mining techniques have significant advantages in processing large-scale data and identifying potential patterns. For example, Wang Qin and Wang Qian [9] developed computer user identification technology by using cluster analysis technology and related algorithms, and realized the user identity verification function through preprocessing, pattern mining and cluster analysis of Internet user log information. To accurately calculate the consumption behavior of college students and analyze the impact of environmental factors on it, Huang [10] used data mining methods to cluster and analyze the consumption behavior data. According to the principles of data mining, she extracted data cluster centers, performed data mining and dataset transformation, determined the main consumption patterns, and combined the collected data samples to solve clustering parameters and statistical vectors. However, these methods have shown a lack of flexibility in capturing dynamically changing user behavior patterns. Therefore, this article proposes a mining algorithm based on the combination of ensemble learning and time series analysis, aiming to solve the dynamic nature of user consumption behavior and data noise problems, thereby improving the accuracy and real-time performance of pattern recognition.

3 Methods

3.1 Data Preprocessing and Quality Assurance

Data preprocessing and quality assurance are crucial steps before conducting consumer behavior pattern recognition. Raw transaction data often contains noise, missing values, and outliers. If analyzed directly, it can easily lead to bias or even misleading model results. Therefore, the core goal of data preprocessing is to improve the integrity, consistency, and availability of data, thereby providing a reliable foundation for subsequent analysis.

Firstly, the noise and invalid information in the original data were cleaned. Noise data mainly manifests as duplicate transaction records, entries with inconsistent formats, and obviously invalid entries (such as negative transaction amounts or abnormal timestamps). This article cleans up by removing completely duplicate records and using simple rules to determine invalid entries (Table 1).

For example, records with timestamp anomalies, such as transaction times displayed as future dates or entries that clearly exceed a reasonable time range, will be deleted. In

Table 1. Data issues and handling methods

Data Issue	Handling Method	Remarks
Missing Values	Mean Imputation/Interpolation	Suitable for numerical data
Outlier Detection	IQR Method/Z-score Method	Based on predefined thresholds
Format Inconsistency	Standardization using Regular Expressions	Dates, category codes, etc

addition, regular expressions are used to unify string data of different formats, such as dates, product category codes, etc., to ensure consistency in data format.

Missing values are a common issue in transaction data. For numerical data such as consumption amount, transaction frequency, etc., this article adopts various imputation strategies. For small missing values, mean or median imputation methods were used to reduce the impact of the imputation process on the overall data distribution. For large missing fields, interpolation or K-nearest neighbor algorithm based on similar transaction records are used to fill them in, ensuring the rationality of the data after filling. For certain fields that cannot be reasonably filled, missing records were directly deleted to ensure that they do not cause bias in the analysis results [11].

When processing consumer data, the presence of outliers may reflect abnormal user behavior or data entry errors. To ensure the accuracy of the analysis, this article uses two main methods for outlier detection and processing. Firstly, for numerical features such as consumption amount, transaction frequency, etc., this article uses the quartile range method and Z-score (standard score) method to detect outliers. When the absolute Z-score of a certain data exceeds the set threshold, it is considered an outlier and the context determines whether to correct or delete it.

In addition, for discrete variables such as product category, payment method, etc., by analyzing frequency distribution, extremely rare or non business logical entries are filtered out. For example, if a certain product category has only appeared once or twice in a specific consumer group, the data may belong to abnormal input rather than the user's actual consumption behavior.

Abnormal inputs will not have a negative impact on subsequent clustering and classification models. This article normalizes numerical data. Specifically, numerical features such as consumption amount and transaction frequency are normalized by Min Max to map the data to the interval of [0,1], keeping each feature at the same level. This step can ensure that the weights between features can be balanced in subsequent modeling, avoiding excessive influence of high-value features on the results.

Finally, this article conducted consistency checks on the processed data to ensure that the logical relationships between the data are correct and error free. Especially for time series data, the order of transactions and whether the changes in consumption patterns of the same user during different time periods conform to common sense were verified. In addition, the uniqueness of users was also checked to ensure consistency between each user's ID and their transaction records, preventing errors or confusion in user transaction data due to data merging or processing issues.

3.2 Extraction of Consumer Behavior Characteristics

Consumption frequency is an important indicator for measuring user activity, which directly reflects users' consumption habits. To extract this feature, this article counted the number of transactions per user during a specific time period (such as month, quarter, etc.). For user groups with obvious cyclical consumption behavior, consumption frequency can help this article further understand their consumption patterns [12].

The consumption amount is another key feature that can reflect the user's purchasing power and consumption preferences. In the process of feature extraction, this article not only calculated the total consumption amount of users, but also separately calculated their average consumption amount, maximum single consumption amount, and minimum single consumption amount.

The types of products purchased by users can directly reflect their consumption preferences and lifestyle. To extract this feature, this article calculated the consumption frequency and proportion of consumption amount of each user in different product categories. Specifically, calculating the total consumption amount and transaction frequency of users in each category according to the category of goods (such as electronic products, clothing, food, etc.).

In addition, the diversity of product types is also an important feature. By calculating the number of product types purchased by each user, this article can further distinguish between "wide net type" consumer users and "precision purchase type" consumer users. For example, the more types of products purchased by users, the more diverse their consumption needs may be, while users who purchase relatively single types may have strong preferences for a certain type of product.

The characteristics of purchase time can help this article analyze users' consumption cycles and time preferences. In the process of feature extraction, this article recorded in detail the purchase time of each user, and extracted the distribution characteristics of the user's transaction time period (such as morning, afternoon, evening) and purchase cycle (such as daily, weekend, holiday). Through these data, this article can analyze whether users have specific consumption time preferences.

3.3 Cluster Analysis of Consumer Behavior Among User Groups

In the analysis of user consumption behavior, there are significant differences in consumption habits among different user groups. Therefore, using clustering methods to classify users reasonably is a key step in in-depth analysis of consumption behavior patterns. Cluster analysis can not only help identify the consumption characteristics of user groups in this article, but also provide important basis for personalized marketing and precise recommendations. In this study, the K-means clustering algorithm was used to cluster and analyze users' consumption behavior, and the optimal number of clusters was selected through reasonable indicators.

K-means is a commonly used unsupervised learning algorithm that can allocate sample data to different clusters, where samples within each cluster have high similarity in certain features. This algorithm is implemented by minimizing the objective function

of the following formula (1):

$$J = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

In formula (1), K represents the number of clusters, C_i represents the i -th cluster, x represents sample points, and μ_i represents the centroid of the i -th cluster. This algorithm iteratively updates the centroids of clusters, assigns samples to the clusters corresponding to their nearest centroids, and ultimately achieves the maximization of intra cluster sample similarity and inter cluster sample difference. According to the Euclidean distance formula, assigning each sample to the nearest cluster, as shown in formula (2):

$$d(x, \mu_i) = \sqrt{(x_1 - \mu_{i1})^2 + (x_2 - \mu_{i2})^2 + \dots + (x_n - \mu_{in})^2} \quad (2)$$

In this study, the multidimensional consumption characteristics of users (such as consumption frequency, consumption amount, product types, etc.) were selected as input data for K-means clustering analysis.

A key issue in cluster analysis is to determine the optimal number of clusters K . In this study, the optimal K value was determined using the “Elbow Method”. Specifically, when running the K-means algorithm in this article, the K value is gradually increased, and the Within Cluster Sum of Squares (WCSS) is calculated for different K values. When the K value is small, the intra cluster error will significantly decrease with the increase of K , but when the K value reaches a critical point, the decrease in error becomes less significant, forming an “elbow” shape. The K value corresponding to the “elbow” point is considered the optimal number of clusters.

3.4 Association Rule Mining and Consumption Pattern Analysis

Association rule mining is one of the core technologies in the field of data mining, which excels in revealing hidden patterns and associations in datasets, especially in analyzing user consumption behavior. By mining frequent itemsets in user consumption data, potential connections between consumers’ common preferences, product combinations, and purchasing behaviors can be discovered. This study used the Apriori algorithm to systematically analyze user transaction data, identify typical consumption patterns, and provide key data support for precision marketing and product recommendation [13].

Apriori algorithm is a classic association rule mining algorithm, mainly used to discover frequent itemsets from large-scale transaction data. This algorithm gradually expands the size of frequent itemsets to mine association rules that meet the set support and confidence thresholds. In this study, each record in the transaction data represents a user’s consumption behavior, with goods as the transaction item. By applying the Apriori algorithm, it is possible to reveal the product combinations that users often purchase together during a single shopping trip, as well as whether the purchase of certain products will affect the purchase decisions of other products.

In the specific implementation process, this article first conducted preliminary processing on the user’s transaction records, treating the goods in each transaction as a set and inputting it into the Apriori algorithm. By setting certain support and confidence

thresholds, the algorithm outputs multiple frequent itemsets and association rules. The calculation formula for support and confidence can be expressed as formula (3–4):

$$\text{Support}(A \Rightarrow B) = \frac{\text{Transactions containing A and B}}{\text{Total transactions}} \quad (3)$$

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Transactions containing A and B}}{\text{Transactions containing A}} \quad (4)$$

Support reflects the frequency of a product combination appearing in all transaction records, while confidence represents the probability of another product being purchased when one product is purchased.

Through further analysis of the mining results, this article identified some representative consumption patterns and behavioral characteristics:

High correlation product combinations: Some product combinations have high support and confidence, such as electronic products and accessories, sports equipment and clothing, food and daily necessities. These product combinations are usually items that users frequently purchase in the same consumption. By analyzing these combinations, it can be inferred that users' consumption preferences, such as purchasing related accessories at the same time as purchasing electronic products, indicate that users have higher additional demands when consuming electronic goods.

Seasonal consumption patterns: By mining association rules on consumption records from different time periods, it was found that the purchasing behavior of some products has obvious seasonality. For example, in summer, users are more inclined to purchase cool clothing and beverages, while in winter, they prefer warm products and hot drinks. This type of seasonal pattern can provide reference for businesses to optimize inventory and marketing strategies at different times.

Linkage promotion opportunities: Association rule mining also reveals potential linkage promotion opportunities. For example, some users have a 70% chance of purchasing sportswear at the same time as purchasing sports shoes, which provides opportunities for joint promotions for businesses. By recommending relevant sportswear to users when purchasing sports shoes, overall sales can be effectively increased. This consumption behavior pattern provides direct data support for personalized recommendations and product matching sales.

4 Results and Discussion

4.1 Clustering Effect Evaluation Experiment

In the clustering effect evaluation experiment, the K-means clustering algorithm was evaluated for its clustering performance on simulated user consumption data. In the experiment, simulated data with three types of features were generated, and different K values were used for clustering analysis to calculate the average contour coefficient for each K value. The experiment analyzed the clustering performance under different K values by drawing a line graph of K value and contour coefficient, as shown in Fig. 1:

In this experiment, K-means clustering analysis was performed on simulated user consumption data, using different K values (2 to 10) for clustering and calculating the

average silhouette coefficient. The results show that when K is around 4, the average contour coefficient reaches its maximum value, with a specific value of 0.73, indicating that the clustering scheme has the best effect. This is consistent with the actual number of categories in the generated data, indicating that the K value can effectively distinguish user groups. As the K value further increases, the contour coefficient gradually decreases, indicating that excessive clustering will lead to a decrease in clustering effectiveness. Therefore, $K = 4$ is the optimal choice for this clustering.

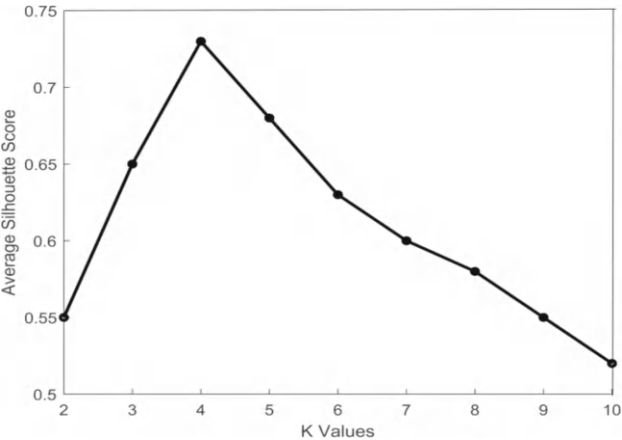


Fig. 1. Clustering effect evaluation

4.2 Evaluation Experiment on the Effectiveness of Association Rule Mining

In the experiment of evaluating the effectiveness of association rule mining, the influence of different support and confidence thresholds on the number of rules generated in association rule mining was evaluated. In the experiment, transaction data from 1000 users were tested and different support and confidence thresholds were set. In the experiment, the support threshold was set from 0.05 to 0.5, and the confidence threshold was set from 0.5 to 0.9. The relationship between the two and the number of rules was plotted separately, as shown in Fig. 2:

The impact of different support and confidence thresholds on association rule generation was evaluated through experiments. The results show that as the support threshold increases from 0.05 to 0.5, the number of generated rules gradually decreases from 400 to 40; when the confidence threshold increases from 0.5 to 0.9, the number of rules decreases from 600 to 333. This indicates that higher support and confidence thresholds will significantly reduce the number of generated rules, and the rules mined will be more concentrated on high-frequency and highly correlated product combinations. Therefore, reasonable selection of support and confidence thresholds can effectively control the number of rules and optimize the mining effect of association rules.

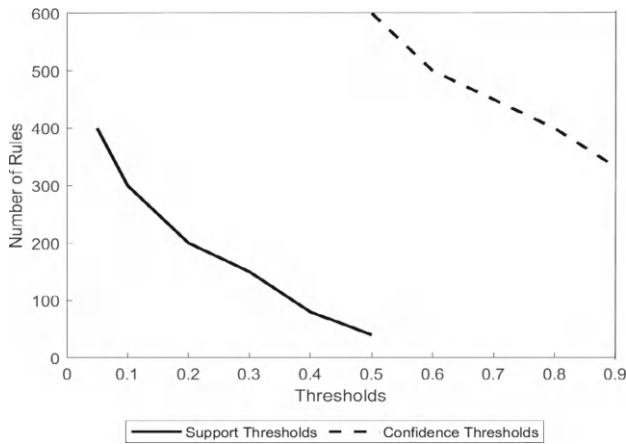


Fig. 2. Evaluation of association rule mining effectiveness

4.3 Time Series Prediction Accuracy Evaluation Experiment

In the time series prediction accuracy evaluation experiment, the ARIMA (1,1,1) model was used to predict simulated user consumption behavior time series data. Firstly, generating time series data with trends and seasonality to simulate real changes in consumer behavior. Next, training the ARIMA model on the data from the first 100 time points and predict the values for the next 20 time points. Finally, MSE and MAPE were used to evaluate the predicted results, and the actual values were visually compared with the predicted values, as shown in Fig. 3:

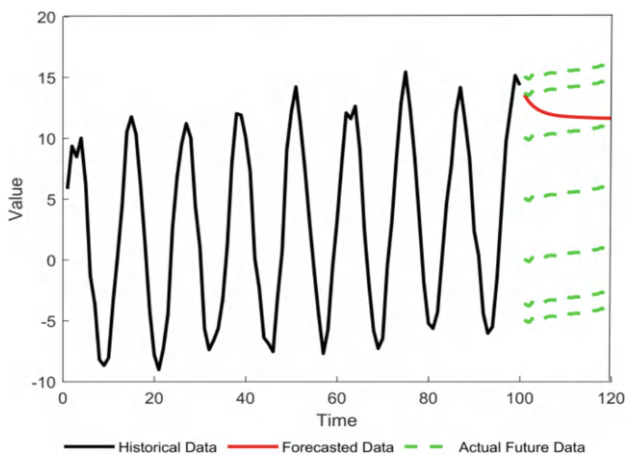


Fig. 3. Machining accuracy evaluation

In this experiment, the ARIMA (1,1,1) model was used to predict the simulated time series data. The results show that the MSE between the predicted and actual values for

the next 20 time points is 0.8421, and the MAPE is 7.63%. From the figure, it can be seen that the model captures the trend and seasonal fluctuations of the data well, and the error between the predicted values and the actual values is within an acceptable range. This indicates that the ARIMA model has high prediction accuracy when processing time series data with trends and seasonality, and is suitable for similar consumer behavior prediction tasks.

4.4 Classification Prediction Accuracy Evaluation Experiment

The classification prediction accuracy evaluation experiment assessed the accuracy of the random forest classification model in multi class classification tasks. In the experiment, a random forest model was used for training and predictions were made on the test set. The experimental results were evaluated through confusion matrix and various classification performance indicators, as shown in Table 2:

Table 2. Evaluation of classification prediction accuracy

Metric	Class 1	Class 2	Class 3	Average
Precision	0.75	0.8	0.7	0.75
Recall	0.7	0.85	0.65	0.73
F1 Score	0.72	0.82	0.67	0.74
Accuracy	-	-	-	0.77

According to the experimental results, the average classification accuracy of the random forest model on the test set is 0.77. The average precision of each category is 0.75, the recall is 0.73, and the F1 value is 0.74. The predictive performance of the model varies slightly across different categories, with category 2 showing better accuracy and recall, while categories 1 and 3 have relatively lower performance. However, overall, the model performs stably and is suitable for multi class classification tasks.

5 Conclusion

This article uses data mining techniques to systematically analyze and mine users’ transaction data, successfully identifying their consumption behavior patterns. In the specific implementation process, the data was first cleaned and feature extracted, and then the K-means algorithm was used to cluster users and identify the characteristics of different consumer groups. Subsequently, the Apriori algorithm is used to mine association rules between users and analyze common product combinations and consumption preferences. In addition, the ARIMA model was used to predict users’ consumption trends in time series, capturing the cyclical changes in consumption behavior. Finally, a random forest classification model was used to predict future consumer behavior and achieved satisfactory results. Although the method proposed in this article can effectively analyze

and predict user consumption behavior, there are still some shortcomings. For example, when dealing with extremely large datasets, the computational efficiency of algorithms still needs to be improved. In addition, when dealing with more complex changes in consumer behavior, the model may need to introduce more advanced machine learning algorithms. Future research can combine deep learning techniques to further improve the accuracy and robustness of models, while optimizing algorithm performance to adapt to larger scale transaction data.

References

1. Anitha, P., Patil, M.M.: RFM model for customer purchase behavior using K-Means algorithm. *J. King Saud Univ.-Comput. Inform. Sci.* **34**(5), 1785–1792 (2022)
2. Aziz, N., Aftab, S.: Data mining framework for nutrition ranking: methodology: SPSS modeller. *Int. J. Technol. Innov. Manag.* **1**(1), 85–95 (2021). <https://doi.org/10.54489/ijtim.v1i1.16>
3. Ageed, Z.S., Zeebaree, S.R.M., Sadeeq, M.M., et al.: Comprehensive survey of big data mining approaches in cloud systems. *Qubahan Acad. J.* **1**(2), 29–38 (2021)
4. Moinuddin, M., Usman, M., Khan, R.: Decoding consumer behavior: the role of marketing analytics in driving campaign success. *Int. J. Adv. Eng. Technol. Innov.* **1**(4), 118–141 (2024)
5. Safara, F.: A computational model to predict consumer behaviour during COVID-19 pandemic. *Comput. Econ.* **59**(4), 1525–1538 (2022)
6. Santoso, M.H.: Application of association rule method using apriori algorithm to find sales patterns case study of indomaret tanjung anom. *Brilliance: Res. Artif. Intell.* **1**(2), 54–66 (2021)
7. Liao, S., Widowati, R., Lee, C.Y.: Data mining analytics investigation on TikTok users' behaviors: social media app development. *Library Hi Tech* **42**(4), 1116–1131 (2024)
8. Saura, J.R.: Using data sciences in digital marketing: framework, methods, and performance metrics. *J. Innov. Knowl.* **6**(2), 92–102 (2021)
9. Wang, Q., Wang, Q.: Computer user behavior analysis and recognition based on data mining. *Software* **44**(5), 139–141 (2023)
10. Meiting, H.: Research on clustering statistics of college students' consumer behavior data based on data mining. *Inform. Comput.* **35**(4), 4–6 (2023)
11. Liu, J., Yu, Y., Mehraliyev, F., et al.: What affects the online ratings of restaurant consumers: a research perspective on text-mining big data analysis. *Int. J. Contemp. Hosp. Manag.* **34**(10), 3607–3633 (2022)
12. El Aouifi, H., El Hajji, M., Es-Saady, Y., et al.: Predicting learner' s performance through video sequences viewing behavior analysis using educational data-mining. *Educ. Inf. Technol.* **26**(5), 5799–5814 (2021)
13. Cherif, A., Badhib, A., Ammar, H., et al.: Credit card fraud detection in the era of disruptive technologies: a systematic review. *J. King Saud Univ.-Comput. Inform. Sci.* **35**(1), 145–174 (2023)



Hierarchical Scheduling Method of Power Emergency Based on Differential Evolution Algorithm

Kuiwen Huang¹(✉), Taiping Yuan², Haowen Yu¹, Jie Zhu¹, and Huajun Tang²

¹ Guangxi Power Grid Company Guilin Power Supply Bureau, Guilin, China
lunwen2024@aliyun.com

² CSG Energy Development Research Institute, Guangzhou, China

Abstract. In the emergency management of power system, efficient material dispatching is of great significance for rapid response to emergencies. In view of the challenges of traditional scheduling methods in complex multi-objective optimization, this paper explores a multi-objective hierarchical scheduling method based on differential evolution algorithm. The upper layer focuses on the timeliness and economy of resource scheduling, while the lower layer focuses on the fairness and coverage of material allocation. The results show that the optimized algorithm performs well in inverse generation distance, dispersion, scheduling fairness and resource coverage satisfaction, which verifies the adaptability and stability of the algorithm. This method provides a reference for power enterprises to make efficient decisions in emergency management, and helps to reduce the risk of power system and improve the reliability of power supply.

Keywords: Hierarchical Scheduling Method · Power Emergency · Differential Evolution Algorithm

1 Introduction

In the event of an emergency in the power system, the ability to quickly and effectively dispatch materials is crucial for handling emergencies and maintaining stable power supply. However, traditional material scheduling methods often find it difficult to achieve both speed and economy while ensuring fairness when facing situations where multiple objectives need to be considered simultaneously [1]. With the increasing scale and complexity of the power system, how to achieve efficient and reasonable material scheduling in emergency situations has become an urgent problem that needs to be solved. We propose a new multi-objective hierarchical scheduling method based on differential evolution algorithm. We hope to provide some assistance to power companies in making efficient decisions in emergency situations by optimizing resource allocation and material distribution.

Differential evolution (DE) is a population-based evolutionary algorithm proposed by Rainer Storn and Kenneth Price in 1995. It is a global optimization algorithm suitable

for handling a variety of optimization problems, including nonlinear, multimodal, and high-dimensional problems. Differential evolution algorithm performs well in solving complex optimization problems, mainly because its principle is simple and based on real coding design, which makes the operation process easy to understand and implement, and reduces the threshold of use. Different from some algorithms that need complex parameter adjustment, the differential evolution algorithm only needs to set two parameters: the scaling factor and the crossover rate, which simplifies the adjustment process. At the same time, the algorithm supports a variety of mutation strategies, which can flexibly adapt to different problems and improve the overall performance [2]. In addition, the algorithm has a significant advantage in robustness, is not sensitive to the initial value, and has a high fault tolerance in parameter settings, even if the parameters are biased, it can still maintain stable operation. These characteristics make it have strong adaptability and stability in multi-objective optimization.

2 Multi-objective Hierarchical Scheduling Method of Power Emergency Based on Differential Evolution Algorithm

2.1 Framework of Multi-objective Hierarchical Scheduling Method

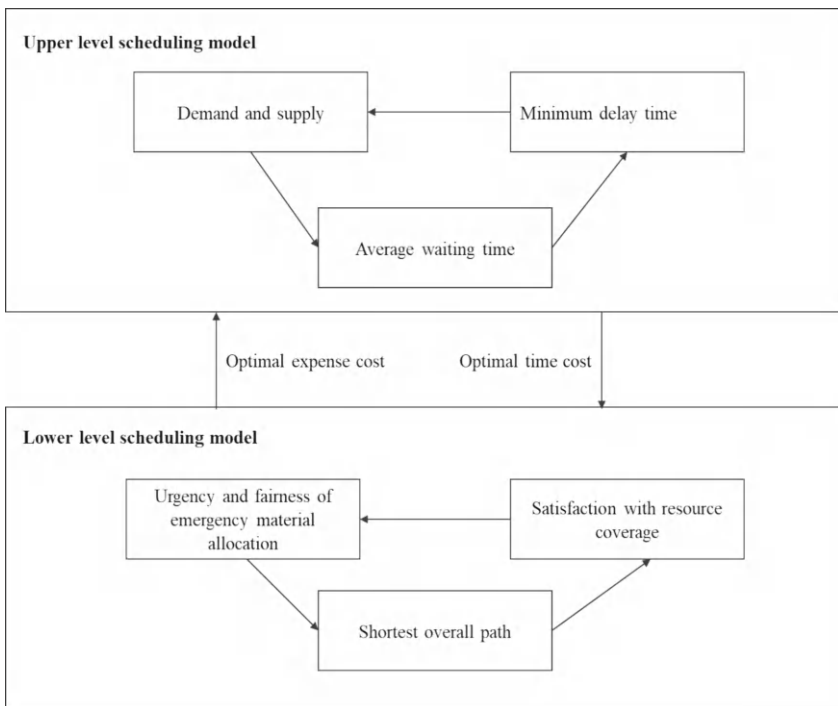


Fig. 1. Multi-objective hierarchical scheduling method architecture for emergency supplies

The multi-objective hierarchical scheduling method for emergency supplies is based on a two-tier scheduling model, which aims to cope with the complexity and diverse needs of resource allocation in emergency situations (Fig. 1). The core of upper scheduling is to optimize the efficiency and economy of resource scheduling in order to achieve rapid and low-cost material transportation [3]. Specifically, by shortening the scheduling delay time, it ensures that emergency supplies can quickly arrive at the fault site to meet the emergency needs in emergencies. Through reasonable transportation route arrangement and tool selection, the dispatching cost is optimized, the overall cost is reduced, and the goal of economy and efficiency is achieved. The upper layer scheduling also focuses on reducing the average waiting time to improve the efficiency of the overall resource allocation with refined scheduling plans and dynamic adjustments.

The lower-level scheduling focuses on the fairness and coverage of material distribution. In case of emergency, the demand intensity in different areas is different, so the lower-level dispatcher gives priority to the most urgent location through urgency analysis, and ensures the rational allocation of resources among the demand points, so as to avoid the impact of uneven distribution on rescue effectiveness. In addition, the satisfaction of resource coverage is one of the key indicators of the lower-level scheduling, that is, to ensure that each demand point can obtain sufficient materials, so as to achieve higher allocation satisfaction [4].

In the aspect of transportation path optimization, the lower scheduling model emphasizes the design of the shortest total path to reduce transportation time and cost and avoid unnecessary waste of resources. This strategy can not only speed up the supply allocation, but also improve the overall utilization efficiency in the case of resource constraints. Through the design of two-tier architecture, the scheduling system can achieve efficient configuration and rational allocation of resources in emergency situations, and provide reliable solutions for emergency management.

2.2 Objective Function of Multi-objective Hierarchical Scheduling

In the process of emergency material scheduling, in order to improve the efficiency of resource allocation and the timeliness of emergency response, it is necessary to build a multi-objective hierarchical scheduling optimization model. The model is divided into two levels, the upper level and the lower level, and the objective functions of each level are from different perspectives to ensure the efficiency of resource scheduling and the rationality of allocation, and ultimately achieve the comprehensive optimization of emergency management [5].

In the upper-level scheduling, the main goal is to control the speed of resource transportation and the economic cost. Firstly, in order to reduce the delay time of resource scheduling in emergency response, set T_i For the first i The optimization objective function of the total delay time can be expressed as: $\text{Minimize } Z_1 = \sum_{i=1}^n T_i$.

Among, n Represents the total number of tasks. By minimizing the delay time of each task, the transportation progress of materials and the start-up speed of emergency repair work can be effectively accelerated. In addition, it is necessary to optimize the economic cost in the dispatching process. Set C_i For the first i The scheduling cost (including transportation and storage cost) of tasks, and the optimization objective is: $\text{Minimize } Z_2 = \sum_{i=1}^n C_i$.

The cost can be effectively controlled on the premise of ensuring timeliness through reasonable selection of transportation routes and allocation of resources. In addition to these two aspects, it is also necessary to minimize the average waiting time at each demand point. Set W_i is the waiting time of i -th demand point, and the optimization objective can be expressed as: Minimize $Z_3 = \frac{1}{n} \sum_{i=1}^n W_i$.

The significance of shortening the average waiting time is to enable each demand point to obtain the required materials in the shortest possible time, thus greatly improving the overall emergency response efficiency.

The lower scheduling objective focuses on the fairness of material distribution and the maximization of coverage. First of all, the urgency and fairness of emergency supplies need to be considered comprehensively. Set U_i and F_i respectively represent for the urgency and fairness scores of each demand point, the optimization objective of the lower layer scheduling can be expressed as: Maximize $Z_4 = \sum_{i=1}^m (\alpha U_i + \beta F_i)$.

Among the formula, m represents the total number of demand points; α and β are the weight coefficients to balance the relationship between urgency and fairness, to ensure that resources are allocated to urgent demand points first, and to ensure the fairness of allocation. Further considering the degree of demand for materials at each demand point, the satisfaction of resource coverage has also become a key optimization objective. Set S_i is the satisfaction of the i -th demand point, and its optimization objective is: Maximize $Z_5 = \sum_{i=1}^m S_i$.

In order to improve the overall efficiency of resource allocation, it is also important to reduce the total length of transportation paths. Set L_k is the length of k -th transportation route, and the optimization objective of the total path length is: Minimize $Z_6 = \sum_{k=1}^p L_k$.

Through the above multi-objective optimization model, the two-level scheduling system can be optimized comprehensively from a global perspective, which not only speeds up the efficiency of material transportation and distribution, but also ensures the rationality and economy of resource allocation. The flexible control of priority is realized by adjusting the weight coefficient among the objective functions, so as to effectively deal with the diversified emergency demand scenarios.

2.3 Function Solution of Double-Layer Scheduling Objective

In the double-level scheduling optimization of emergency supplies, it is very important to choose the appropriate algorithm to solve the objective function. Based on the complexity of the scheduling model and the requirement of multi-objective optimization, differential evolution algorithm becomes the first choice because of its flexibility and global search ability. Like the solution process shown in Fig. 2, this algorithm can effectively deal with continuous variable optimization problems, and has a high convergence rate and diversity of solutions, so it is suitable for solving multi-objective functions in emergency material scheduling.

In the application process of differential evolution algorithm, the core steps include population initialization, mutation, crossover and selection. In order to improve the global search ability and avoid falling into local optimal solution, the design of mutation strategy and crossover operator is optimized. Specifically, the dynamic mutation strategy

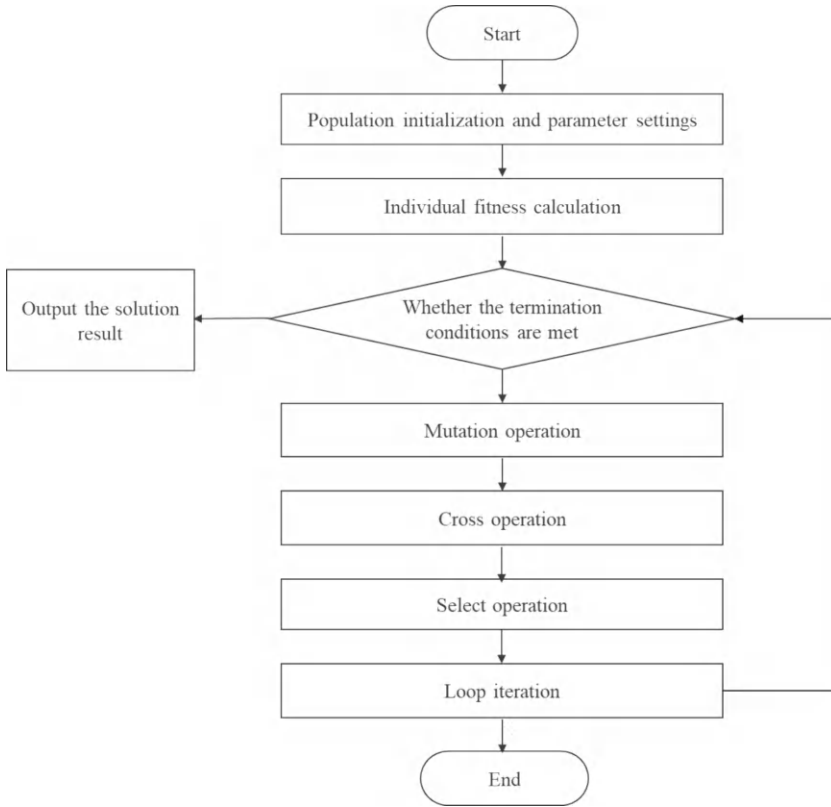


Fig. 2. Solving process of objective function based on differential evolution algorithm

based on the evolution process information is introduced, and the algorithm can maintain good solution exploration ability in different optimization stages by dynamically adjusting the mutation probability [6].

In the process of mutation, the distance between the individual and the optimal solution in the population is taken as the basis for adjusting the mutation probability. Let the i -th individual of the current population is X_i , while the optimal solution is X_{best} . When the distance between individual X_i and X_{best} is large, it means that the individual is far away from the global optimal solution, and the mutation probability is increased at this time. Encourage individuals to explore new search spaces; On the contrary, if the distance is small, the mutation probability is reduced to maintain the stability of the solution. In addition, the mutation rate can also be dynamically adjusted according to the number of successive failures of individuals in the optimization process, so that the convergence process is smoother. The adjustment of the mutation operator is expressed as:

$$p_{mut} = p_0 + \gamma \cdot \left(\frac{d(X_i, X_{best})}{d_{max}} \right) + \delta \cdot F_{fail} \quad (1)$$

Among the formula, p_0 is the initial mutation probability; γ and δ are the adjustment coefficients, $d(X_i, X_{best})$ denotes the distance between the individual and the optimal solution; d_{max} is the largest distance in the population; F_{fail} is the number of consecutive failures.

The crossover operator is optimized by dynamically adjusting the mutation probability, and the crossover probability of each individual is associated with its current fitness to ensure that the diversity is promoted by a higher crossover probability in the early exploration stage, and the accuracy of the solution is improved by gradually reducing the crossover probability in the later convergence stage of the algorithm. Set C_i is the fitness of the i -th individual, the crossover probability is adjusted as follows:

$$p_{cross} = p_{base} \cdot \left(1 - \frac{C_i - C_{min}}{C_{max} - C_{min}} \right) \quad (2)$$

Among, p_{base} Is the base crossover probability is the base crossover probability, C_{min} And C_{max} The minimum and maximum fitness values in the population, respectively. This design can effectively maintain the diversity of solutions and accelerate the convergence speed of the algorithm in the whole differential evolution process, so as to ensure that the bilevel scheduling optimization problem of emergency supplies can be effectively solved in a reasonable time [7]. By constantly adjusting the mutation and crossover operators, the differential evolution algorithm can explore the solution space more comprehensively, and finally find a set of optimal scheduling schemes.

3 Test and Analysis

3.1 Experimental Environment and Data

This experiment was conducted on the MATLAB software platform. The data we use is obtained from a power company, which contains a lot of important information about emergency material needs. These pieces of information include the geographical location of the places where supplies are needed, how much supplies are needed, where supplies are available, and how much supplies these supply points can provide [8]. The location information of supply points not only helps us calculate the distance between them and demand points, but also affects transportation costs and efficiency. The quantity of materials that supply points can provide tells us how much they can support at most. By analyzing these data together, our algorithm can allocate resources reasonably, ensuring that material allocation is both effective and efficient. In the experiment, we also made detailed allocations based on the demand for different types of materials, so that we can develop better allocation strategies for each type of material. In MATLAB, we used a method called differential evolution algorithm to find the best emergency material allocation plan through multiple attempts and calculations. Then we evaluated this plan to ensure that our approach is reliable and efficient [9].

3.2 Algorithm Parameter Setting

Setting algorithm parameters is crucial. The detailed settings of algorithm parameters are shown in Table 1. We adjusted the crossover rate to 0.52, which determines the

degree of genetic information mixing between newly generated individuals and their ‘parents’. It also directly affects whether the algorithm can find diverse solutions and how fast it converges. We set the mutation rate to 0.01. It ensures that our solution does not become too singular during the optimization process, avoiding us from only finding a local optimal solution rather than a global one. We also set the maximum number of evolutionary generations to 500, so that the algorithm will continue to try within this range until the best solution is found. Meanwhile, we also adjusted the individual learning factor to 1.25 and the inertia weight to 0.5.

Table 1. Detailed setting table of algorithm parameters

Parameter	Value	Explain
Crossover rate	0.52	Determine the crossover ratio of genetic information between the newly generated individual and the parent individual
Mutation rate	0.01	Regulate the frequency of mutation operation to maintain the diversity of the population
Maximum evolution algebra	500	Number of algorithm iterations to ensure sufficient search time
Individual learning factor	1.25	Affect the learning ability of the individual in the search space
Inertia weight	0.5	Control the inertia of the individual in the search process and influence the search direction

In the aspect of mutation probability optimization, a dynamic adjustment strategy is adopted to improve the adaptability of the algorithm, and the mutation probability calculation formula is as follows:

$$\xi_i(t) = \begin{cases} \xi_i(t-1) \cdot \phi & \text{if } f_i(t) < f_m(t) \\ \xi_i(t-1) \cdot (\xi_{min} + \xi_{max} - \xi_{min}) \div \eta(t) & \text{if } f_i(t) \geq f_m(t) \end{cases} \quad (3)$$

In this formula, $\xi_i(t)$ indicate the crossover probability of the t -th generation of the i -th individual. ξ_{max} and ξ_{min} represent the maximum and minimum value of the crossover rate respectively. Fitness value $f_i(t)$ and $f_m(t)$ are used to judge the relative performance of individuals and populations, so as to dynamically adjust the crossover probability and ensure the flexibility of the algorithm in global search.

The calculation of mutation rate also adopts the dynamic adjustment strategy, and its calculation formula is as follows:

$$\kappa(t) = \kappa_{min} + (\kappa_{max} - \kappa_{min}) \cos\left(\frac{\pi t}{T}\right) \quad (4)$$

This formula combines the current iteration count with the total iteration count. As we iterate, it will gradually reduce the mutation rate. The advantage of doing so is that it can make the optimization process more stable and avoid large fluctuations in the later stages.

4 Evaluation Indicators and Test Results

When evaluating algorithm performance, we mainly used two indicators: anti generational distance and dispersion. These two indicators can help us understand the effectiveness of algorithms in handling multi-objective optimization problems. The anti-generation distance is an indicator used to evaluate the concentration of the algorithm's solution results [10]. The smaller the value, the wider the distribution of solutions, which usually means that the algorithm can better explore the entire search space and find more different solutions. The indicator of dispersion reflects the diversity of solutions. The smaller the value, the more concentrated the solution found, which may mean that the algorithm is too focused on a certain area and may miss out on good solutions in other areas.

For multi-objective optimization problems, maintaining diversity of solutions and strong global search capabilities are key to finding good solutions. We compared the algorithm performance before and after adjusting the crossover probability. The experimental results show that the optimized algorithm has significant improvements in both the anti-generation distance and dispersion metrics. This indicates that the optimized algorithm not only converges faster, but also maintains the diversity of solutions during the process of finding solutions. The specific data are shown in Table 2.

Table 2. Test results of inverse generation distance and scatter

Solution scale (unit)	Inverse generation distance		Dispersion	
	Optimize	Not optimized	Optimize	Not optimized
3	0.012	0.034	0.15	0.37
6	0.018	0.029	0.18	0.41
9	0.016	0.035	0.20	0.50
12	0.022	0.038	0.12	0.39
15	0.019	0.032	0.16	0.44
18	0.025	0.033	0.14	0.35
21	0.028	0.047	0.13	0.38
24	0.023	0.042	0.21	0.40
27	0.030	0.048	0.17	0.37
30	0.020	0.036	0.26	0.46

It can be seen from Table 2 that under all solution set sizes, the optimized inverse generation distance is lower than that without optimization, which shows that the optimized crossover probability strategy enhances the global search ability of the algorithm, makes the solution results more dispersed, and avoids premature convergence. At the same time, the reduction of the scatter index shows that the algorithm can maintain a certain degree of diversity among different solutions in the case of optimization, so as to

avoid falling into local optimal solution. Overall, the optimized algorithm shows higher flexibility and adaptability.

In addition, the scheduling effect of the algorithm is also evaluated, mainly examining the fairness of the scheduling results and the satisfaction of resource coverage. The fairness index reflects the equilibrium degree of each demand point in the allocation of emergency supplies, while the satisfaction of resource coverage measures the adequacy and coverage of supplies.

Table 3. Fairness of scheduling results and resource coverage satisfaction test results

Fault demand point	Fairness		Satisfaction of resource coverage	
	Optimize	Not optimized	Optimize	Not optimized
1	0.942	0.871	0.940	0.825
2	0.953	0.844	0.957	0.792
3	0.964	0.822	0.963	0.891
4	0.980	0.826	0.982	0.893
5	0.930	0.902	0.978	0.801
6	0.935	0.870	0.956	0.880
7	0.954	0.892	0.962	0.825
8	0.970	0.895	0.980	0.879
9	0.987	0.897	0.987	0.850
10	0.950	0.916	0.972	0.875

It can be seen from Table 3 that the optimized scheduling results are better than unoptimized conditions in terms of fairness and resource coverage satisfaction. After optimization, the average value of fairness has increased by about 8%, indicating that the balance of each demand point in material allocation has been improved, which is helpful to achieve more reasonable material allocation in emergency situations. In addition, the satisfaction of resource coverage is also significantly improved, indicating that more demand points can obtain sufficient material supply under the action of the optimization algorithm, thus enhancing the efficiency and effectiveness of the overall scheduling.

5 Conclusions

A multi-objective hierarchical scheduling method for power emergency based on differential evolution algorithm is explored to deal with the complex optimization challenges in emergency material scheduling in the paper. By designing a two-level scheduling model, the upper level optimizes the timeliness and economy of resource scheduling, while the lower level focuses on the fairness of material allocation and the balance of coverage. The experimental results show that the proposed method performs well in terms of inverse generation distance, dispersion, scheduling fairness and resource coverage

satisfaction, which verifies the adaptability and stability of the algorithm. In addition, the evaluation of risk consolidation function before and after dispatching shows that the proposed method has certain potential in reducing the risk of power system and improving the reliability of power supply.

References

1. Yu, X., Li, C., Zhou, J.F.: A constrained differential evolution algorithm to solve UAV path planning in disaster scenarios. *Knowl.-Based Syst.* **204**, 106209 (2020)
2. Guvenc, U., Duman, S., Kahraman, H.T., et al.: Fitness-Distance Balance based adaptive guided differential evolution algorithm for security-constrained optimal power flow problem incorporating renewable energy sources. *Appl. Soft Comput.* **108**, 107421 (2021)
3. Liu, D., Hu, Z., Su, Q., et al.: A niching differential evolution algorithm for the large-scale combined heat and power economic dispatch problem. *Appl. Soft Comput.* **113**, 108017 (2021)
4. Chen, X.: Novel dual-population adaptive differential evolution algorithm for large-scale multi-fuel economic dispatch with valve-point effects. *Energy* **203**, 117874 (2020)
5. Deng, W., Xu, J., Song, Y., et al.: Differential evolution algorithm with wavelet basis function and optimal mutation strategy for complex optimization problem. *Appl. Soft Comput.* **100**, 106724 (2021)
6. Kiani, E., Doagou-Mojarrad, H., Razmi, H.: Multi-objective optimal power flow considering voltage stability index and emergency demand response program. *Electr. Eng.* **102**, 2493–2508 (2020)
7. Gong, Y., Liu, P., Liu, Y., et al.: Robust operation interval of a large-scale hydro-photovoltaic power system to cope with emergencies. *Appl. Energy* **290**, 116612 (2021)
8. Lu, K.D., Wu, Z.G.: Constrained-differential-evolution-based stealthy sparse cyber-attack and countermeasure in an AC smart grid. *IEEE Trans. Industr. Inf.* **18**(8), 5275–5285 (2021)
9. Wan, Y., Zhong, Y., Ma, A., et al.: An accurate UAV 3-D path planning method for disaster emergency response based on an improved multiobjective swarm intelligence algorithm. *IEEE Trans. Cybern.* **53**(4), 2658–2671 (2022)
10. Qureshi, T.N., Javaid, N., Almogren, A., et al.: An adaptive enhanced differential evolution strategies for topology robustness in internet of things. *Int. J. Web Grid Serv.* **18**(1), 1–33 (2022)



Construction of Movie Knowledge Graph and Design of Recommendation System Based on Movielens Dataset Expansion

Peng Dong^(✉)

School of Business, Stevens Institute of Technology, Hoboken, NJ 7030, USA
pengdong0128@gmail.com

Abstract. As internet technologies continue to advance, users are increasingly overwhelmed by excessive information, making it crucial to have effective systems for personalized content suggestions. However, when new users or insufficient data are present, traditional recommendation systems can suffer from issues like limited data and the cold start problem, which can degrade the quality of recommendations. In this research, a new movie knowledge graph model is proposed using the Movielens-1M dataset, alongside an adaptive recommendation method that combines graph-based learning and attention mechanisms. The approach improves the performance and interpretability of the knowledge graph by refining entity similarity calculations and optimizing the simplification of connection paths. It also strengthens the recommendation system's capacity to perform well even with minimal data. Results from the experiments indicate that the proposed method excels in both prediction accuracy and user satisfaction, particularly in overcoming challenges related to data insufficiency and initial user interactions.

Keywords: Recommendation System · Graph Convolutional Network · Data Sparsity · Cold Start Problem · Movielens Dataset

1 Introduction

Users often find it difficult to effectively obtain the required content when facing massive data. This phenomenon greatly increases the difficulty of information screening. In order to solve this problem, intelligent recommendation systems have gradually become an important means to optimize user experience. By deeply analyzing user behavior patterns and interest preferences, the system can help users filter out potential content of interest, thereby improving usage efficiency. However, although traditional recommendation algorithms have played a role to a certain extent, they are highly dependent on historical behavior data. When user interaction information is scarce or new users have just joined, traditional algorithms often cannot accurately capture user needs, especially in data sparseness and cold start situations. The performance is particularly unsatisfactory.

By connecting entity nodes of multiple dimensions and attributes, knowledge graphs enable recommendation systems to more comprehensively explore users' potential interests and inject rich semantic information into recommendation algorithms. In the field of

movie recommendations, knowledge graphs, with their high flexibility, help the system provide more personalized recommendation services based on users' viewing habits. At the same time, by deeply mining the multiple correlations between movies, the diversity and accuracy of recommendation results are significantly improved. However, although knowledge graphs provide new ideas for improving recommendation systems, existing systems still face data sparsity and cold start problems, and there are many limitations in the structure construction and optimization path of knowledge graphs, which directly affect the accuracy and efficiency of recommendations.

2 Related Research

In recent years, recommendation systems have been widely used to help users make decisions from massive data, especially in the field of movie recommendation [1]. Traditional movie recommendation systems usually rely on two basic methods: collaborative filtering and content filtering, which generate personalized recommendations by analyzing user behavior and preferences. However, with the increasing complexity of application scenarios, traditional methods have gradually shown their limitations in solving problems such as cold start, data sparsity, and malicious attacks. Therefore, many innovative methods have been proposed to improve the efficiency and accuracy of recommendation systems.

TT Ajith proposed a movie recommendation system that combines knowledge graph and particle filtering technology [2]. By running directly in the database, it mines user preferences (such as director, type, etc.) and provides more personalized movie recommendations. Compared with traditional systems based on machine learning and clustering algorithms, this method has significant improvements in efficiency and accuracy. SS Choudhury proposed a matrix measurement method based on trust propagation [3], combining user similarity with trust propagation, and successfully solved the problems of cold start, data sparsity and malicious attacks in recommendation. Experimental results show that the deep neural network (DNN) model combined with trust propagation has better recommendation effect.

In terms of deep learning applications, S Aramuthakannan designed a movie recommendation system based on Taymon optimized deep learning network (TODL net) [4], which combines dilated convolutional neural network (DiCNN) with bidirectional LSTM to overcome the problem of information loss and optimize the recommendation effect by combining user behavior patterns and movie content. The movie recommendation accuracy of this model is as high as 97.24%. P Mondal proposed a recommendation system that combines user feedback (such as like, dislike, neutral) and audio and video information of movie trailers [5]. This method further improves the accuracy of recommendations, especially in the application of Indian language movie datasets.

The widespread application of recommendation systems plays an important role in the user selection process, especially in the context of the surge in the number of similar products, which puts users under greater pressure to choose. The recommendation system analyzes user preferences and provides them with the best products, significantly improving decision-making efficiency. In this regard, KK Jena used artificial neural network to build a movie recommendation system [6], and improved the prediction

accuracy by optimizing the neuron weights. Finally, through the Hit-Ratio evaluation, the model achieved a recommendation accuracy of 87%. The UISVD++ model proposed by D Lliu successfully alleviated the cold start problem and improved the prediction accuracy of the recommendation system by integrating movie types and user age attributes into the SVD++ framework [7].

In order to further improve the effect of traditional recommendation systems, BY Madhavi proposed a hybrid method based on content filtering, which generates a personalized movie recommendation list by integrating multiple text-to-vector conversion technologies and algorithm results [8]. This method breaks through the limitations of traditional single text vector conversion and similarity calculation methods, and significantly improves the diversity and accuracy of recommendations. Z Yuan's research proposed a hybrid recommendation system based on weighted classification and user collaborative filtering algorithm [9], combining sparse linear models with local recommendation models based on user clustering, and further improving the recommendation accuracy by converting item category preferences into low-dimensional dense matrices.

R Lavanya recommends movies through implicit feedback (such as ratings), solving the data sparsity problem [10]. Using the logistic regression (LR) algorithm, the system achieved 81.9%, 69.82% and 32.5% accuracy, precision and recall respectively, verifying the effectiveness of this method in movie recommendation. With the continuous development of technology, future movie recommendation systems will be more intelligent and personalized, and can provide users with more accurate and efficient recommendation services.

3 Construction and Optimization of Knowledge Graph

3.1 Knowledge Graph Construction Method Based on MovieLens Dataset

The dataset contains rich user rating information, various attributes of movies (such as name, type, director, actor, etc.) and other related data, providing a solid foundation for the construction of the knowledge graph. During the construction process, by extracting key entities in the data, such as movies, users, directors, actors, etc., and combining them with related attributes, a multi-level knowledge graph can be formed. Movie nodes form a complete movie information network by establishing connections with multiple attribute nodes such as type, director, and actor. User nodes are connected to movie nodes through behavioral data such as ratings and viewing time to capture user interests and preferences. Through these relationships, the recommendation system can understand user needs more accurately and provide personalized recommendations.

When dealing with data sparsity problems, algorithms such as collaborative filtering and matrix decomposition can be used to optimize the graph structure and fill in the missing rating information. At the same time, combined with natural language processing technology, the text information of the movie (such as descriptions, comments, etc.) is analyzed to extract valuable sentiment and keyword data from it, further enhancing the expression ability of the graph. Through these technical means, the accuracy of the movie knowledge graph has been improved, thereby better supporting the personalized recommendation of the recommendation system.

Graph neural network can effectively capture the potential similarities between movies and changes in user preferences by aggregating the adjacent nodes of each node in the graph. The recommendation system can not only solve the problems of data sparsity and cold start, but also continuously adjust the recommended content as user behavior changes to achieve more accurate personalized recommendations. Combining the text information in the Movielens dataset with the graph data, and extracting the characteristics and emotional tendencies of the movie through natural language processing technology, will further enrich the graph information and improve the recommendation effect. Building a knowledge graph based on the Movielens dataset not only provides a structured data representation for the recommendation system, but also provides strong support for improving the intelligence level of the recommendation system.

3.2 Embedded System Design and Visualization Implementation of Knowledge Graph

The movie knowledge graph integrates and graphs various types of information about the movie into a hierarchical database, so that the semantic information of the movie can be accurately extracted and expressed in a simplified dimension. Each movie usually has multiple key features, including title, director, actor, movie type, language, release time, country of production, and rating, etc. These features are often interrelated and closely connected.

In the user set part, it contains the unique identifier of each user and related personal information, including basic information such as gender and age. In the movie set part, detailed information such as the ID, name, type and so on of each movie is provided to distinguish and filter various types of movies. The dataset contains more than 1 million rating records, ranging from 1 to 5 points, and divided into nine rating levels with a span of 0.5 points, which makes the distribution of rating data more detailed. The distribution of the number of movies at each rating level in the data, as shown in Fig. 1, can provide detailed support for further analysis and the construction of recommendation systems.

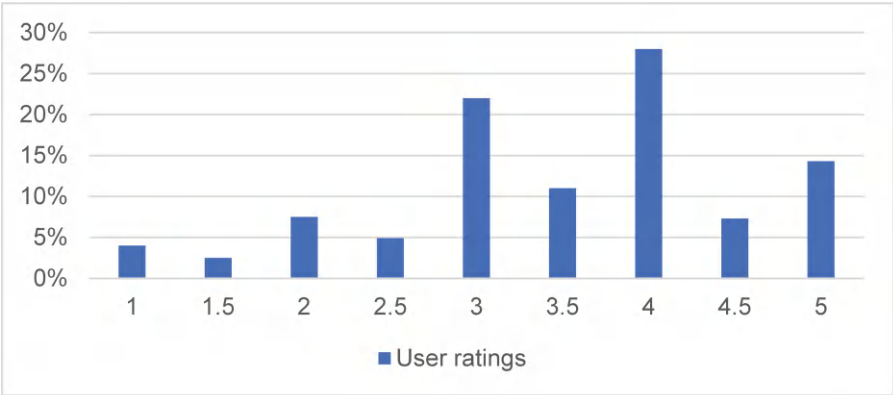


Fig. 1. User rating statistics

In the process of building a movie knowledge graph, we first store information from multiple data sources in the data layer, and then process it through fine screening and standardization to make it meet the requirements of the model layer. This process involves integrating structured, semi-structured and unstructured data with external third-party databases, and finally generating a high-quality knowledge graph through data cleaning and knowledge processing. Structured data is directly stored in the graph after standardization, semi-structured data is obtained through information extracted from web pages, and unstructured data includes pictures, audio and video related to movies. The advantage of graph databases is that they can reduce the dimension of complex high-dimensional semantics, intuitively express data relationships, improve query efficiency and facilitate practical application operations.

Through the Neo4j graph database, we successfully built a movie knowledge graph, which stores and manages the basic attributes of movies and the relationships between them in the form of triples. Each movie and its related information are organized in order in the form of nodes and relationships, thereby improving data query efficiency and access speed. The system not only supports users to query detailed information about movies, but also recommends related movies based on the associated data stored in the graph, helping users discover more movies of interest. To simplify the graph structure, we removed redundant paths to ensure that the graph is clearer and more concise, and improve the accuracy of queries. With the continuous addition of new movie data, the knowledge graph has been effectively expanded, enhancing the diversity and accuracy of recommendations, laying the foundation for the optimization of personalized recommendation systems and the application of future algorithms.

The architecture of a movie recommendation system usually consists of multiple modules, including user interaction interfaces, recommendation engines, and display pages. During the operation of the system, the user's behavior data will first be collected and analyzed, and then processed through a series of specific algorithm models to generate a movie recommendation list that meets user needs. When using the system, users can personalize their accounts by registering them. At the same time, users can also update their personal information at any time, and use the collection, rating, and comment functions provided by the system to share their views and comments on movies. In addition, users can also share their favorite movies with other users with the same interests through social functions. In order to ensure that the recommendation system can provide accurate recommendation results, the system is highly dependent on user behavior data.

4 Design and Experimental Evaluation of Adaptive Recommendation Algorithm

4.1 Multi-layer Adaptive Recommendation Model Based on Graph Convolutional Network

In order to further improve the recommendation effect, we design an adaptive similarity learning method to automatically discover and match user groups similar to the target user by analyzing user behavior data. In view of the sparse user rating data, we adopt an

improved similarity measurement method, combined with cosine similarity and modified cosine similarity, and optimize the similarity calculation process through dynamic weighting strategy, avoiding the deviation caused by data sparsity in traditional methods. Finally, through adaptive feature learning, we can provide accurate movie recommendations for new users and users with scarce data, significantly improving the performance and user experience of the recommendation system.

The system gathers user behavior data and interacts with movie information to build a connection model between users and films. By implementing a graph-based approach, the system dynamically updates user profiles and recommended movie lists, aiming to boost the precision of recommendations. This method identifies hidden preferences by assessing the correlations between user features and movie attributes, and learns node representations within a knowledge graph through graph-based techniques. Through successive layers of aggregation and expansion, the system can better capture the complex relationships between user tastes and movie characteristics, thereby improving the quality of suggestions.

In this approach, information is shared between adjacent nodes, considering both individual node attributes and the attributes of connected nodes, allowing for continuous refinement of recommendations. This iterative process addresses issues like cold start and enhances both the precision and diversity of the recommendations. The key to the recommendation strategy lies in progressively updating the information at each node, refining the user's interest profile, and generating personalized suggestions via an optimization mechanism. To ensure the accuracy of the suggestions, the model is trained using a loss function that minimizes the difference between actual user interactions and predicted values, as outlined in formula (1).

$$L = - \sum_{k=1}^K [y_{ui} \log(\hat{y}_{ui}) + (1 - y_{ui}) \log(1 - \hat{y}_{ui})] \quad (1)$$

Among them, y_{ui} is the actual interaction value between the user and the movie, \hat{y}_{ui} is the predicted interaction probability, and K is the number of interaction instances.

Additionally, to further strengthen the model's capability to represent various relationships in the knowledge graph, this work integrates the TransE framework for training entity triplets. By optimizing the associated loss function of these triplets, the effectiveness of the graph-based convolutional model is improved. The formula for the loss function of the triplets is given in Eq. (2). In this case, $f(h,r,t)$ quantifies the distance between the head entity, the relationship, and the tail entity, while σ signifies the sigmoid activation function. By leveraging the combined strengths of graph convolutional approaches and the TransE model, the proposed recommendation system in this study is able to more accurately capture user preferences, providing highly personalized and precise recommendations.

$$L = \sum_{(h,r,t) \in G} \log(\sigma(f(h, r, t)) - f(h, r, t')) \quad (2)$$

4.2 Experimental Results and Analysis

In the recommendation system, the Top-N recommendation model generates a recommendation list to evaluate whether it meets user needs. In practical applications, the

recommendation effect is evaluated by checking whether the product that the user interacts with appears in the recommendation list. Precision and recall are commonly used evaluation indicators. Precision measures the proportion of correct recommendations in the recommendation list, reflecting the accuracy of the recommendation system; recall measures the proportion of user-interested products that the recommendation system can find among all relevant products, reflecting the recall rate.

In addition to precision and recall, the ranking quality of the recommendation system is also very important. In order to evaluate the ranking effect, the normalized discounted cumulative gain (NDCG) indicator is used. NDCG comprehensively considers the ranking of products in the recommendation list, and higher rankings lead to higher scores. DCG first calculates the gain of each recommended product and then performs a discount based on its ranking. To evaluate the effectiveness of the recommendation results, the NDCG score is derived by comparing it to the discounted cumulative gain of the ideal ranking. A holistic performance assessment of the recommendation model is then made by considering precision, recall, and NDCG together.

This study is carried out on a system running the Ubuntu distribution of Linux, utilizing an NVIDIA GeForce RTX 3090 graphics card for hardware acceleration. The AKGCN model, along with other competing algorithms, is executed using the PyTorch library. The specifics of the experimental setup are detailed in Table 1 below.

Table 1. Experimental environment configuration data

The name of the parameter	Set the value	The name of the parameter	Set the value
Number of samples per batch	1024	Regularization coefficient	1e−5
Training rounds	100	Neighbor samples	4
Training rounds	100	Neighbor samples	4

In this study, the training process of the AKGCN model relies on a graph database built with Neo4j. In this process, unique identifiers are first assigned to each user and movie, and then the relationship between users and movies is constructed based on these identifiers, and these relationships are used to construct the adjacency matrix of similar entities. By extracting relevant entities and their interconnected relationships from the knowledge graph and aggregating these data, the model can obtain more accurate entity representations and relationship patterns. Subsequently, the processed movie dataset is loaded into the model for initialization, and the relevant training parameters are set, and the Top-N recommendation index is selected to evaluate the recommendation effect of the model.

The training process starts with random shuffling of the dataset, and then the model is trained with different amounts of data. Through multiple iterations of calculation, the model can gradually adjust its parameters to achieve optimization. In this process, some data that cannot be fully classified will be discarded to ensure the integrity and validity of the training data. Finally, the model’s precision, recall, and NDCG evaluation

indicators will be used to comprehensively measure the performance of the recommendation system to ensure that the model can provide efficient, accurate, and personalized recommendation results that meet actual application needs.

This paper assesses the effectiveness of several recommendation algorithms using two list lengths, 20 and 100, and applies precision ($P@k$), recall ($R@k$), and normalized discounted cumulative gain (NDCG) as performance indicators. The outcomes of the experiments are summarized in Table 2, Figs. 2, and 3, with the specific data presented below.

Table 2. Comparison of normalized discounted cumulative gain of recommendation algorithms

Algorithm	Data set	N@20	N@100
MKR	Movielens-1M	0.0684	0.1366
NFM	Movielens-1M	0.0575	0.1292
KGCN	Movielens-1M	0.0704	0.1416
AKGCN	Movielens-1M	0.0754	0.1563
CKE	Movielens-1M	0.0610	0.1326

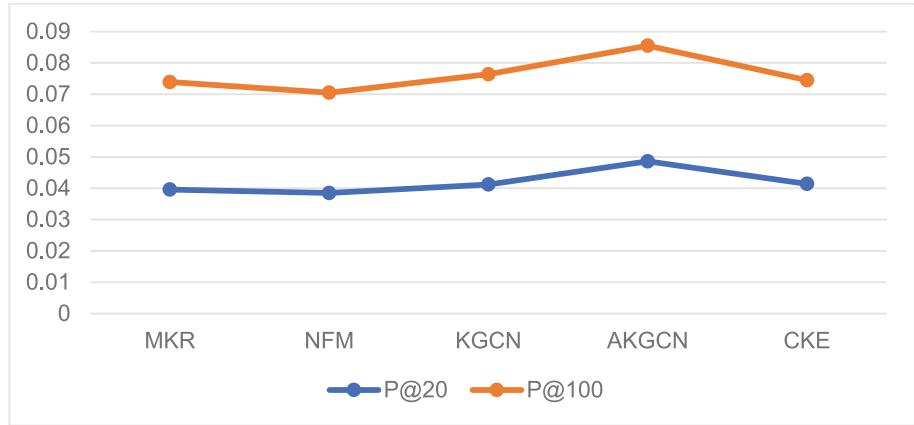


Fig. 2. Comparison of the accuracy of recommendation algorithms

From the experimental results, in the test of Movielens-1M dataset, the AKGCN model proposed in this paper showed superior performance in all indicators. Compared with other recommendation algorithms, especially KGCN model, AKGCN’s recommendation effect is significantly improved. Its core advantage is that by introducing adaptive attention mechanism, it accurately captures the implicit association between users and movies, so as to make the recommendation results more accurate.

Through detailed analysis of these results, it can be found that NFM model performs the worst, and its precision and recall indicators are lower than other models. The performance of CKE and MKR models is at an intermediate level. CKE model is slightly

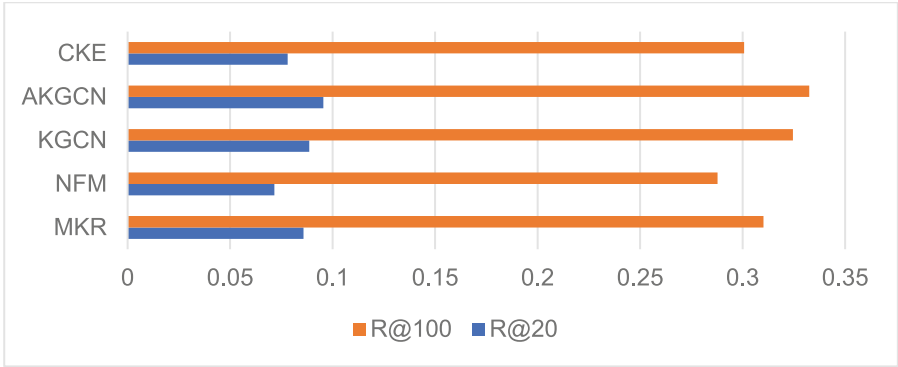


Fig. 3. Comparison of the recall rate of recommendation algorithms

better than MKR at P@20, but its performance in other evaluation criteria is not satisfactory. The specific reason is that CKE model has difficulties in processing semantic information and cannot effectively mine useful features in data. MKR and KGCN models can better extract semantic information based on knowledge graph, so their overall performance is better. In particular, KGCN model significantly improves the accuracy of recommendation system by aggregating multi-layer adjacency information. However, KGCN needs to process huge graph data during training, which increases the complexity of calculation and is easily disturbed by noisy data.

Expanding upon this concept, the AKGCN model introduces an attention mechanism to facilitate adaptive learning, optimizing feature extraction and effectively filtering out irrelevant data, ensuring that the recommendations are more aligned with user interests. This method proves particularly advantageous in overcoming issues like data insufficiency and cold starts, allowing AKGCN to generate more precise recommendations under these conditions. The experimental results show considerable improvements across various evaluation metrics, reinforcing the model’s practical potential in recommendation systems.

The findings indicate that the accuracy of all models improves as the number of training cycles increases. However, models like MKR and KGCN experience some inconsistency in their accuracy. For recommendation lists of length 20, the CKE, MER, and KGCN models show comparable accuracy, with CKE marginally outperforming MKR. However, when the recommendation list length is extended to 100, MKR surpasses CKE in terms of accuracy. As the list grows, particularly in the Movielens-1M dataset, the recommendations become less precise, as some suggested movies do not align with the user’s preferences, leading to a decrease in accuracy. Despite this, the AKGCN model consistently delivers the best results in all test settings, showing superior performance in both 20-item and 100-item lists.

In terms of recall, all models show notable improvements as training progresses, with a more pronounced effect when the list length is 20. As the length of the recommendation list increases to 100, recall stabilizes across models, and a significant enhancement in

recall is observed, with values rising from below 0.1 to nearly 0.3. Among the models, NFM exhibits the lowest recall, while CKE, MKR, and KGCN perform similarly. However, AKGCN stands out by achieving the highest recall across all configurations.

When evaluating NDCG, significant differences appear when the recommendation list length is 20, but the performance of CKE and MKR models converges as the list length increases to 100. Throughout the training process, the NDCG score of AKGCN continues to increase, with a particularly notable rise when the recommendation list reaches 100 items, consistently outshining the other models.

By embedding the attention mechanism into the model, AKGCN enhances the power of graph convolutional networks, enabling more accurate alignment with user preferences. This improvement boosts both precision and recall, demonstrating AKGCN's advantage in the realm of recommendation systems.

5 Conclusion and Outlook

With the advent of the information age, information overload has become a major problem in user decision-making, and personalized recommendation systems have played an important role in solving this problem. The recommendation system based on knowledge graph proposed in this paper successfully improves the recommendation performance by introducing the adaptive knowledge graph convolutional network (AKGCN) model. Experiments have shown that the system performs well in indicators such as precision, recall and NDCG, proving the effectiveness of combining knowledge graphs with graph convolutional neural networks. However, the current model still has problems such as limited knowledge graph content and inaccurate prediction of user preference changes. Future research should further expand the richness of knowledge graphs, optimize the structure of graph convolutional neural networks, and pay attention to the dynamic changes of user behavior to improve the personalization and adaptability of the system. In addition, with the development of big data technology, the ability of recommendation systems in large-scale data processing and storage needs to be continuously enhanced, and future optimization directions are still full of potential.

References

1. Ivaturi, S.S.R., Thalatham, M.N.V., Bugatha, A.S.K., et al.: Hybrid movie recommendation system based on user preferences and item similarity. In: Lin, F.M., Patel, A., Kesswani, N., Sambana, B. (eds.) *Accelerating Discoveries in Data Science and Artificial Intelligence I: ICDSAI 2023*, LIET Vizianagaram, India, April 24–25, pp. 671–681. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-51167-7_64
2. Ajith, T.T., Ajay, K.C.V., Nandakishore, J., et al.: Enhanced movie recommendation using knowledge graph and particle filtering. *IEEE* (2021). <https://doi.org/10.1109/ICOCSE51865.2021.9591834>
3. Choudhury, S.S., Mohanty, S.N., Jagadev, A.K.: Multimodal trust based recommender system with machine learning approaches for movie recommendation. *Int. J. Inf. Technol.* **13**(2), 475–482 (2021). <https://doi.org/10.1007/s41870-020-00553-2>
4. Ziaee, S.S., Rahmani, H., Nazari, M.: MoRGH: movie recommender system using GNNs on heterogeneous graphs. *Knowl. Inf. Syst.* **66**, 7419–7435 (2024)

5. Mondal, P., Kapoor, P., Singh, S., et al.: Task-specific and graph convolutional network based multi-modal movie recommendation system in Indian setting. *Procedia Comput. Sci.* **222**, 591–600 (2023). <https://doi.org/10.1016/j.procs.2023.08.197>
6. Jena, K.K., Bhoi, S.K., Mallick, C., et al.: Neural model based collaborative filtering for movie recommendation system. *Int. J. Inf. Technol.* **14**(4), 2067–2077 (2022). <https://doi.org/10.1007/s41870-022-00858-4>
7. Liu, Y., Miyazaki, J.: Knowledge-aware attentional neural network for review-based movie recommendation with explanations. *Neural Comput. Applic.* **35**, 2717–2735 (2023)
8. Madhavi, B.Y., Monika, D.R., Sunil, S.K.: Movie recommendation system using content-based filtering. *IJARIIIE* **2021**(4)
9. Huang, G., Zhu, X., Wasti, S.H., et al.: Multi-knowledge resources-based semantic similarity models with application for movie recommender system. *Artif. Intell. Rev.* **56**(Suppl 2), 2151–2182 (2023)
10. Lavanya, R., Bharathi, B.: Movie recommendation system to solve data sparsity using collaborative filtering approach. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **20**(5), 1–14 (2021). <https://doi.org/10.1145/3459091>



Design and Implementation of a High Concurrency Online Payment Platform Based on Distributed Microservice Architecture

Tianyou Huang^(✉)

AWS Payments Processing, Amazon, Seattle, WA 98101, USA
richardniw2025@gmail.com

Abstract. In recent years, with the acceleration of the domestic financial industry towards the Internet, electronic payment channels have become increasingly rich, and there are great differences in message format, communication protocol and business logic between channels. To reduce payment costs, enhance user experience, and simplify channel management, online payment platforms need to integrate the commonalities and characteristics between channels. However, many small banks still use traditional single node payment systems, which can no longer meet the needs of high concurrency scenarios in Internet finance. It is urgent to develop a new payment system with high concurrency processing capability and stability and reliability. This article is based on a distributed microservice architecture and designs and implements a daytime transaction module in an online payment platform. The system is developed using Java programming language and SpringBoot framework, combined with distributed features to enhance the system's concurrent processing capability and stability. Its technology stack includes MySQL as database support, Linux system as runtime environment, RabbitMQ providing message queue service, Dubbo implementing remote calling framework, Spring scheduled tasks responsible for scheduling tasks, and Maven for project management. The platform's functions include asynchronous notifications, channel routing, timeout queries, signature verification, and error code handling. It supports core business scenarios such as delegated collection and account recharge, providing practical reference for the design of high concurrency payment systems.

Keywords: Java · Springboot · Distributed Architecture · Microservices · High Concurrency Payment System Java · Springboot · Distributed Architecture · Microservices · High Concurrency Payment System

1 Introduction

In recent years, with the rapid popularization of mobile communication technology and Internet finance, China's mobile payment industry has shown a booming trend. Data shows that by the end of 2018, the total number of mobile Internet users in China had reached 817 million, of which 583 million chose to use mobile payment. The number of online payment transactions completed by non bank institutions throughout the

year reached 530.611 billion, with transaction amounts exceeding 20.807 trillion yuan, an increase of 85.05% and 45.23% respectively compared to the previous year. This field not only has a large user base, but also continues to expand in transaction scale, and the application scenarios are gradually becoming more diverse. From shopping and consumption to transportation, to medical payment and public services, mobile payment is deeply integrated into people's daily lives. In addition, payment technology is constantly innovating, from QR code payment to NFC payment, and then to facial recognition payment. Each technological upgrade significantly improves the convenience of payment and promotes the rise of new business models such as unmanned retail. China's mobile payment technology is gradually moving towards internationalization and has been widely applied in countries such as South Korea and Australia. Despite this, many small and medium-sized banks still use outdated single node payment systems. Their performance is not enough to cope with the high concurrency scenario of Internet finance. The system functions are scarce and the operation and maintenance costs are high, so optimization and upgrading are urgently needed. In response to this pain point, this article proposes a design and implementation scheme for an online payment platform based on a distributed microservice architecture, with a focus on developing a daytime trading system module. By adopting a microservice architecture and modularizing the deployment of functional components, the system has achieved key functions such as channel access, intelligent routing, and asynchronous notifications. It supports core businesses such as delegated collection and account recharge, significantly reducing payment costs, improving system stability and operational efficiency, and providing reliable technical support for high concurrency payment scenarios.

2 Related Research

In the development process of Java information management systems, commonly used technologies include Spring framework, Hibernate ORM, Servlet containers (such as Tomcat), Maven build tools, etc. This type of system is typically used to manage and process various types of information, including but not limited to documents, data, user information, etc. In their article [1], L. Wei et al. studied a library management system based on C/S, Eclipse as the development environment, Sqlyog as the database server, and Java language. This system has the characteristics of fast running speed, high security, and strong portability. In the article [2], J. Xue et al. designed and built a comprehensive public security information management platform based on Java technology and analyzed the comprehensive public security business. The platform received good feedback in actual testing, verifying the feasibility of the system. Y Chen et al. proposed a Java based personal health information management system in the article [3]. Based on the Java platform, the overall hierarchical structure of the system was designed. Y Xue et al. introduced some commonly used development tools in the article [4], then analyzed the user functional and data requirements of the system, and analyzed the feasibility of system development from multiple aspects. Based on the B/S mode, a teaching equipment management information system was designed and implemented using Java language, JSP technology, and MySQL database.

SQL Server is a relational database management system (RDBMS) developed by Microsoft, widely used in enterprise level database management and data processing

fields. It provides powerful data management, security, scalability, and performance optimization capabilities, suitable for applications of all sizes and data storage needs. ZL Xu et al. developed a digital system in the article [5] to manage construction information and monitoring data of foundation pits. A dynamic synchronization analysis system was developed based on SQL Server external database, Visio, and Excel software. Y Yang et al. proposed a design of an electrical automatic control water treatment system based on LabVIEW in their article [6]. Using SQL Server as the backend database to achieve basic monitoring functions such as data collection, device control, real-time curve display, historical recording, and fault alarm.

B. Barua and colleagues developed and implemented a cloud-based online travel portal supported by a distributed database system built on a microservices architecture [7]. Their research highlights the application of various data fragmentation techniques, strategies for data allocation, and the use of relational algebra, union operations, and joins to achieve effective data integration and distribution within the system. M. Shamim and his team proposed a microservices-based architecture aimed at enhancing the adaptability and efficiency of an airline reservation system. The architecture incorporates Redis for caching, employs Kafka and RabbitMQ as messaging systems, and integrates MongoDB alongside PostgreSQL for data storage [8]. To boost scalability, the system leverages Docker and Kubernetes, enabling horizontal scaling to accommodate fluctuating demands. X. Chen's study evaluated a distributed aircraft design environment built on microservices and cloud computing [9]. Compared to conventional collaboration methods, the cloud-based approach significantly shortened the time required for design iterations among team members. M. Panahandeh and his team introduced a novel approach to anomaly detection in microservice systems, named ServiceAnomaly [10]. This method integrates distributed tracing with six analytical metrics to construct an annotated directed acyclic graph, effectively capturing the normal operational behavior of the system.

3 Method

3.1 Microservice Architecture Design

The traditional system architecture involves stacking multiple independent systems and continuously adding new features, resulting in a large and complex system with highly coupled functions. Once a certain function fails, it may affect the normal operation of the entire system. The microservice architecture breaks down system functionality into independent services, each of which can be developed and deployed independently, reducing coupling between systems and improving system availability. The microservice architecture processes a single function through small and lightweight services, communicates between services through lightweight mechanisms, supports multiple languages and storage technologies, and reduces the need for centralized management. It not only improves deployment speed and security, but also reduces the risk of system wide crashes caused by functional failures through independent service modules. In high concurrency environments, reasonable splitting and allocation of resources is the key to microservice architecture, ensuring that the system can run efficiently and stably.

The system adopts an architecture design based on distributed microservices, and is divided into seven main modules, namely: 1. Payment gateway; 2. Online acquiring module; 3. Fund exchange module; 4. Channel gateway; 5. Daytime batch processing module; 6. Notification module; 7. Monitoring Center. The specific composition of the system architecture is shown in Fig. 1.

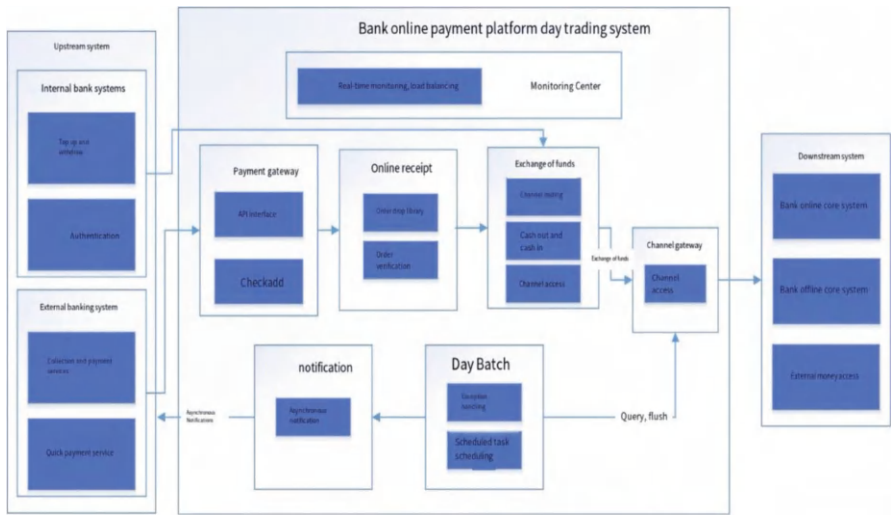


Fig. 1. Architecture diagram of daytime transaction system for bank online payment platform

The system architecture consists of seven functionally independent sub modules, each performing different tasks without interfering with each other. The payment gateway provides external API interfaces to handle external request verification, data conversion, request forwarding, and logging; The online acquiring system is responsible for order verification and data storage; The fund exchange system provides in and out fund services and conducts channel routing selection; The channel gateway is responsible for accessing external payment channels and exposing remote interfaces for other applications to call; The daytime batch processing system provides scheduled tasks and timeout transaction query services; Asynchronous notifications for transaction processing in the notification system; The monitoring center is responsible for real-time monitoring and load balancing of the system. Among these seven systems, the payment gateway, channel gateway, and notification system need to be connected through an external network; The external system can directly access the interface of the payment gateway through the network, while the channel gateway and notification system communicate with the external system. If the transaction originates from the bank’s internal system, it can be processed directly through the fund exchange platform without the involvement of a payment gateway.

Distributed architecture has become the mainstream of system design, mainly divided into two models: symmetric and asymmetric. The asymmetric distributed model deploys components to multiple physical nodes, with service requests starting from the entrance

component’s server and communicating through remote calls. The symmetric distributed model requires all requests to pass through specific physical nodes, which increases the coordination burden. Each node deploys the same data and programs, independently providing services to improve system availability. This article adopts a distributed model based on data partitioning, which divides modules accessing the same database into independent units, avoiding data overlap and improving system performance and scalability.

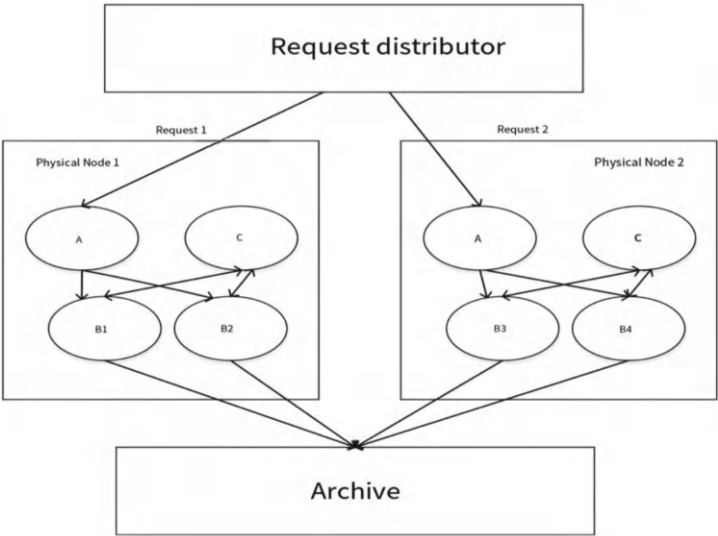


Fig. 2. Distributed Model Based on Data Partition

Figure 2 shows a distributed architecture model based on data partitioning. Modules A and C provide services for core module B, which is responsible for processing data loading related requests. Each data partition sends a request to the request distributor through module A, which then forwards it to the appropriate node. This architecture allocates services based on data partitioning, where service requests are directed to the corresponding data partitioning and distributed across multiple physical nodes to achieve load balancing and avoid single node overload. Interdependent nodes share data partitioning, reduce cross node communication, lower database access frequency, and thus improve performance. Each physical node contains partial data partitioning, supporting horizontal scalability and flexibility of the system.

3.2 Payment Platform Daytime Trading System

(1) Design of timeout query process

When the transaction timeout occurs, the system will store the exception record in the exception registration table and perform scheduled tasks to pull pending transactions,

using multi-threaded parallel processing to handle these exceptions. The system will query downstream transaction results, and if a clear answer is obtained, update the transaction status; If the return result is not clear, the system will check whether the processing times have exceeded the limit. If they have exceeded the limit, the processing will be stopped. If they have not exceeded the limit, the system will continue to wait for the next query. To avoid duplicate processing of tasks, the system will lock transactions and set gradually increasing time intervals to handle each abnormal task.

(2) Asynchronous notification process design

After the transaction timeout occurs, the relevant notification tasks are added to the message queue, and the notification system extracts the tasks from the queue and starts processing them. The system will parse the received notification information, update the merchant's order records, and insert them into the notification registration form. Based on the processing results, the system will notify upstream merchants to ensure the smooth progress of the transaction process.

(3) Design of entrusted collection process

The entrusted collection transaction is initiated by an external merchant, and the payment platform selects the appropriate downstream channel through routing, records the transaction flow, and sends the information. If a transaction timeout occurs, the system will enter the timeout query process and inform the merchant of the final transaction status through an asynchronous notification mechanism. The process includes signature verification, message validation, order recording, fund flow registration, and timeout query to ensure that transactions are completed as expected.

(4) Design of bank account recharge process

Bank account recharge is initiated through internal channels and directly requested to be processed by the fund exchange system. After verifying the request format, the system records the transaction flow and selects the appropriate channel. If the collection transaction fails, a failure message will be returned; If the transaction times out, the system will enter the timeout query process, update the transaction status, and notify the upstream merchant of the final result of the transaction.

4 Results and Discussion

4.1 Analysis of Load Balancing Algorithm

In the data partitioning model, the core of load balancing is to reasonably redistribute data partitioning. Assuming there are n data partitions in the system, each with a database access volume of D . The D value is determined by the number of reads (R) and writes (W) to the database.

In order to consider load changes, two expected parameters TR_i and TW_i were added, representing the changes in read and write operations per unit time. These expected parameters can reflect changes in system load, thereby affecting the calculation of D value, as shown below:

$$D = R + TR_i + V_{\partial} W + TW_i \quad (1)$$

In practical operation, due to the different performance losses of read and write operations, a read and write coefficient V_D is introduced to weight and distinguish their loads. Through these adjustments, the D value can more accurately reflect the load situation of each data partition, thereby achieving dynamic load balancing. In order to avoid service interruption caused by frequent movement of data partitioning, a threshold parameter Dmax is introduced to control when to initiate the redistribution of data partitioning. When the D ratio of a physical node exceeds th, data partitioning migration will be initiated, otherwise the status quo will be maintained. This algorithm arranges the D values of each node in descending order to allocate the load reasonably and optimize the overall performance of the system.

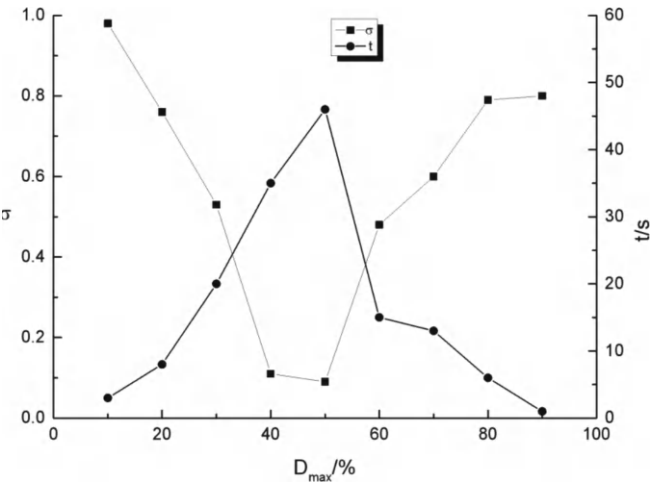


Fig. 3. Dmax impact on σ and t

Figure 3 shows the effect of Dmax value on σ and t . To analyze this relationship, the percentage of the maximum D value of the server is used in the graph to represent Dmax. The observation results show that when the Dmax value is within the range of 10% to 25% of the extreme value of D, the server load can usually easily reach D, making data partitioning difficult to migrate, and at this time, the σ value is relatively large; When the Dmax value increases to 40% to 60%, the σ value decreases while the D value increases significantly; When further increased to 70% to 80%, the D value significantly decreased, while the σ value increased again. If the Dmax value continues to increase, overloading will affect system performance. Therefore, considering the influence of load and t-value, 70% to 80% of the extreme value of Dmax is usually selected as the D value, and this study sets it at 75%.

In this distributed model, VSRT is also one of the key criteria for determining whether data partitioning needs to be reallocated. Assuming Dmax is set to 80% of its maximum value, the impact of VSRT on σ and t is shown in Fig. 4. From the diagram, it can be seen that as the VSRT value gradually increases, the t value continues to decrease, while the σ value gradually increases. However, when VSRT reaches a certain threshold, the

t-value will decrease to zero, and the σ value will no longer increase after reaching a certain range. Therefore, the optimal range for VSRT is 1.4 to 1.6, and this study chose 1.5. Overall, when Dmax is set to 75% of the maximum value of D and VSRT is set to 1.5, the performance of the load balancing algorithm is the most superior.

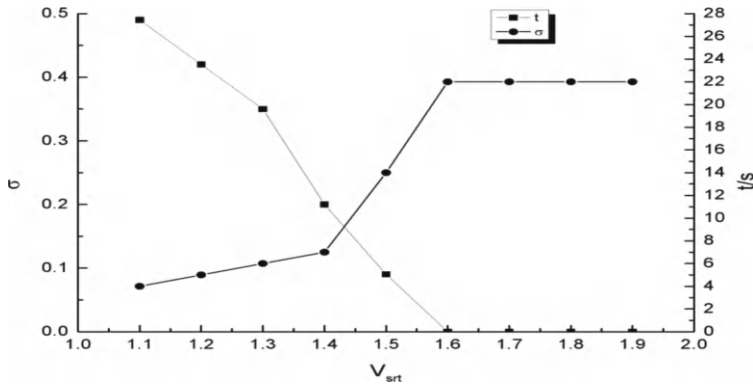


Fig. 4. VSRT impact on σ and t

4.2 Performance Testing

Figure 5 shows the results of single node performance testing, where the x-axis represents concurrent transaction volume, t represents response time, and tps is the average throughput per second. The test results show that with the increase of response time, the response time of both the old and new systems has increased, but overall, the concurrent response time of the old system is higher than that of the new system. When comparing the TPS of the old and new systems, it was found that the throughput of the new system was overall better than that of the old system. However, the TPS of the old system is greatly affected by concurrency in high concurrency scenarios. As concurrency increases, throughput shows a downward trend, reflecting the poor stability of the old system in high concurrency situations. In contrast, the TPS of the new system is more stable and less affected by concurrent access, indicating stronger stability in high concurrency environments.

Next, we will focus on testing in multi node scenarios, comparing the TPS changes of new and old systems under different node numbers in high concurrency scenarios. Figure 6 shows the trend of TPS improvement multiple with the change of concurrent transaction volume under different node numbers for both new and old systems. We tested the changes in system TPS improvement when the number of nodes was increased to 2 and 4, respectively. The results indicate that both the new and old systems have improved TPS with an increase in the number of nodes, demonstrating that both systems have certain load balancing capabilities. However, as the concurrency increases, the TPS improvement factor of the new system remains stable, while the improvement factor of the old system gradually decreases, indicating that the distributed load balancing

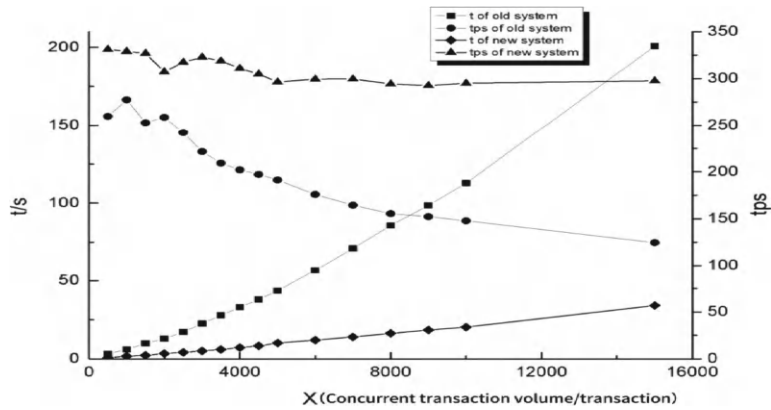


Fig. 5. The variation of system response time/tps with concurrency

algorithm of the old system failed to efficiently distribute the concurrent load to each node, resulting in poor stability. In comparison, the load balancing algorithm of the new system demonstrates better stability in high concurrency situations. In addition, when the number of nodes is 2, the TPS of the new system increases by nearly 2 times, and when the number of nodes is 4, it increases by nearly 4 times. This indicates that the newly designed distributed payment platform can effectively utilize each node and allocate transaction load reasonably in high concurrency scenarios, ensuring the stable operation of the system.

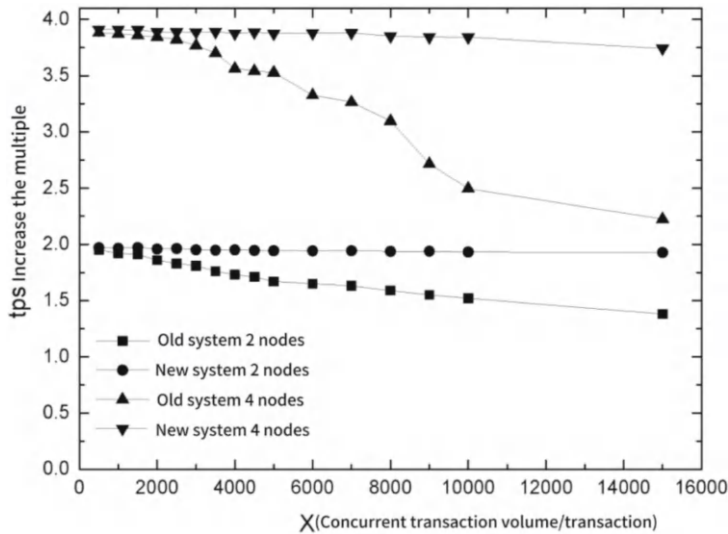


Fig. 6. Changes in TPS Enhancement Ratio with Concurrent Transaction Volume under Different Node Numbers

5 Conclusion

This study designed and implemented a high concurrency online payment platform based on a distributed microservice architecture, aimed at meeting the needs of bank day trading systems. The system adopts technology stacks such as Java, SpringBoot, Maven, Dubbo, etc., and combines a distributed model of data partitioning to build an efficient and stable payment platform. The platform includes multiple modules, such as payment gateway, payment processing, fund processing, connector, scheduling, notification, and monitoring center, which can effectively support the expansion of banking business and enhance user experience. Through in-depth analysis of system functionality and business requirements, a suitable distributed microservice framework was designed, and database, microservice architecture, and corresponding code implementation were provided. In addition, algorithm optimizations such as load balancing, distributed serial number generation, signature verification, channel routing, and error codes ensure the high stability and efficiency of the system. Compared with the original payment system of a certain bank, this system performs better in stability and throughput (TPS). However, there is still room for further optimization of the system, such as introducing a distributed configuration platform to improve configuration management efficiency, and optimizing data reading performance through distributed caching to further enhance the system's performance and user experience in high concurrency environments.

References

1. Zhou, Y., Zhu, Y., Luo, Q., et al.: Optimizing pumped-storage power station operation for boosting power grid absorbability to renewable energy. *Energy Convers. Manag.* **299**, 117827 (2024)
2. Yang, J.: Research on multi-objective peak shaving optimization scheduling of distributed power grids based on grey PSO algorithm. Lanzhou University of Technology (2024)
3. Wu, W.: Research on Distributed Computing and Convex Relaxation Technology for Optimal Energy Flow Problems in Integrated Energy Systems. South China University of Technology (2021)
4. Li, X., Yang, N., Li, Z., et al.: Confidence estimation transformer for long-term renewable energy forecasting in reinforcement learning-based power grid dispatching. *CSEE J. Power and Energy Syst.* **10**(4), 1502–1513 (2024). <https://doi.org/10.17775/CSEEJPES.2022.02050>
5. Zhan, C., Li, X., Lu, Y., et al.: Research on double layered distributed optimization scheduling strategy for active distribution networks based on clusters. *Electr. Transm.* **53**(1), 81–90 (2023)
6. Zhou, Y., Zhong, Y., Yu, D., Tong, J.: Research on economic optimization dispatch of distributed microgrid system based on improved ECA algorithm the electrical era (002) (2022)
7. Barua, B., Whaiduzzaman, M., Mesbahuddin Sarker, M., et al.: Designing and implementing a distributed database for microservices cloud-based online travel portal (2023). https://doi.org/10.1007/978-981-19-5443-6_22.
8. Barua, B., Kaiser, M.S.: Novel Architecture for Distributed Travel Data Integration and Service Provision Using Microservices. [arXiv:2410.24174](https://arxiv.org/abs/2410.24174) (2024)
9. Chen, X., Isoldi, A., Riaz, A., et al.: Evaluation of a collaborative and distributed aircraft design environment, enabled by microservices and cloud computing. In: *AIAA SCITECH 2023 Forum* (2023). <https://doi.org/10.2514/6.2023-1163>
10. Panahandeh, M., Hamou-Lhadj, A., Hamdaqa, M., et al.: ServiceAnomaly: an anomaly detection approach in microservices using distributed traces and profiling metrics. *J. Syst. Softw.* **209**, 111917 (2024). <https://doi.org/10.1016/j.jss.2023.111917>



Design and Implementation of a General Data Collection System Architecture Based on Relational Database Technology

Yuxin Wang(✉)

2970 International Dr. APT 109C, Ypsilanti, MI 48197, USA
yuxinece@gmail.com

Abstract. This study proposes a general data collection system architecture based on relational database technology, aimed at meeting complex and diverse data collection needs. Based on the characteristics of relational databases, the system achieves dynamic correlation and integration of multi-source data through the “knowledge element” model, and designs a universal underlying storage structure that supports multiple collection tasks. By adopting object serialization method, the compatibility issue of different data storage modes has been solved, enabling unstructured data to be stored in a unified JSON string format in relational databases. In addition, based on the logical model of the “self association” table, multi-level forwarding support for deterministic and additive collection tasks has been implemented, and a state transition model between hierarchical tasks has been constructed. In the specific implementation, ASP NET WebForm technology combines transaction scripts and activity recording patterns to dynamically generate data collection interfaces; Implement bidirectional binding between front-end and data through Vue framework, and use Web API to complete dynamic data loading and storage. To solve the common storage problem of tree structured data, the system introduces Level and Path fields, and combines triggers to automatically maintain values, significantly improving the efficiency and automation level of complex data collection tasks.

Keywords: Relational Database · General Data Collection · Task Forwarding · Information System

1 Introduction

In recent years, frequent emergencies have posed a severe challenge to social order and public safety. These events typically have complexity, dynamism, and public impact, and in order to achieve rapid and effective responses, it is necessary to build efficient data collection systems to support scientific decision-making. Data collection tasks typically exhibit characteristics such as time constraints, diverse information types, hierarchical collection and aggregation, and high requirements for data security. At the same time, it is necessary to ensure that data from different sources can be correlated and integrated to meet subsequent processing needs. According to the characteristics of information

collection, common methods include deterministic collection (single entry with fixed format), additive collection (supporting multiple records), and uncertain collection (flexible format and flexible recording). Faced with the requirements of diverse data types and hierarchical tasks, the collection system must have dynamic modeling capabilities, support flexible task distribution and multi-level management, and take into account the security guarantee of data storage and transmission. Although traditional relational databases face certain challenges in dynamic data patterns and multi-level task support, their efficient query capabilities, clear data structures, and optimization mechanisms still give them significant advantages in information system development. Therefore, based on relational database technology, this study designed a universal data collection system architecture that provides an effective solution for rapid response and decision support in emergency scenarios through dynamic modeling, hierarchical task management, and secure storage.

2 Related Research

2.1 B/S Architecture

It refers to a browser/server architecture system, where users access applications through the browser, and the logic and data processing of the application are completed on the server side. This architecture model has better cross platform and convenience compared to traditional C/S architecture. Jin proposed an overall framework design method for a bridge health monitoring system in the article [1]. The interface scheduling of bridge health monitoring is the bus transmission control of the bridge health monitoring system in B/S mode, and the LCD controller is used for analyzing bridge health monitoring information. Y Jiang et al. completed the overall architecture design, functional module division, and database table structure design of the system using browser/server (B/S) architecture and front-end/back-end separation mode in the article [2]. M Ma designed and developed a community property management information system based on B/S mode in the article [3]. The system development adopts a browser/server (B/S) architecture, with Java as the development language and Spring MVC mode as the framework. MySQL database is used, and the source code and database interaction process use the Mybatis framework. FAN Xue wei et al. proposed a remote monitoring system in their article using a new B/S and C/S fusion architecture [4], which unifies the front-end and back-end interaction between B/S and C/S, and achieves server-side sharing. This system has certain scalability and maintainability.

2.2 Dataset Acquisition

R Nagarajan and colleagues developed an automated system called the Traveler Rating Classification System (TRCS). Since the travel review dataset was unlabeled, they utilized the K-means clustering algorithm to divide the data into three clusters [5]. In their study, the decision tree classifier, using the bagging method, achieved an optimal classification accuracy of 97.95%. F Gu and colleagues proposed a geolocation estimation model based on multitask learning (GLML), which combines classification and retrieval

tasks to determine the similarity between query images and images in the dataset [6]. In the study by TT Zhang et al., they combined low-level units with high-level hidden states in LSTM to emulate the attention mechanism of the human perception system. The hierarchical LSTM model they designed is capable of addressing dependencies across different time scales to capture the spatiotemporal correlations of network-level travel times. Additionally, they developed another self-attention module that connects features extracted by LSTM to a fully connected layer, enabling predictions of travel times across all corridors rather than just individual links or routes [7].

2.3 Data Warehousing and Data Mining

AC Jaures In their study [8], used the analysis of the indigenous seasonal climate forecast (ISCF) in Benin employed the travel cost method, descriptive statistics, and the two-step Heckman technique to evaluate its use and economic value. A study by Krishnan et al. [9] applied the Heckman two-step method (1979) and discovered that the European Central Bank (ECB) positively influences firms' outward foreign direct investment (OFDI). The findings revealed that companies with higher leverage and ECB involvement tend to have more significant OFDI. Additionally, MAJi-Liang et al. [10] used the Heckman two-step model to determine the factors influencing family decisions in commercial legume cultivation. They also applied the endogenous treatment regression (ETR) method to assess how commercial legume farming affects family economic welfare.

3 Method

3.1 High-Level System Design

The system has two roles: administrator and regular user, where regular users share some basic permissions with administrators, such as logging into the system, changing passwords, and personal information. The task management of ordinary users includes the management of architecture and meta knowledge, as well as organizational management, while administrators are mainly responsible for the management of organizations and users. To meet the demand for multi-level task forwarding, the system has set up task allocation mechanisms based on different roles. Users can play different roles in the creation, forwarding, and completion of tasks, and achieve a complete data collection process through collaboration. In the specific process, the task publisher creates tasks within the system and assigns them to relevant members. Task members decide whether to forward the task to the next level based on the situation. If forwarding is not necessary, the task completion person will directly submit the task, and the publisher will then conduct data review. If the review fails, the task will be returned for modification. The status of tasks includes nine states: pending release, pending completion, draft, forwarded, submitted, pending review, returned for repair, pending modification, and completed. Due to the multi-level forwarding characteristics of tasks, the states of parent and child tasks usually change independently, but in certain situations, the states of parent and child tasks can affect each other. For example, when the parent task forwards a task, the status of the parent task will change to "forwarded", while the child task will automatically update to "pending completion". This linkage mechanism ensures the efficiency and consistency of task management. As shown in Fig. 1.

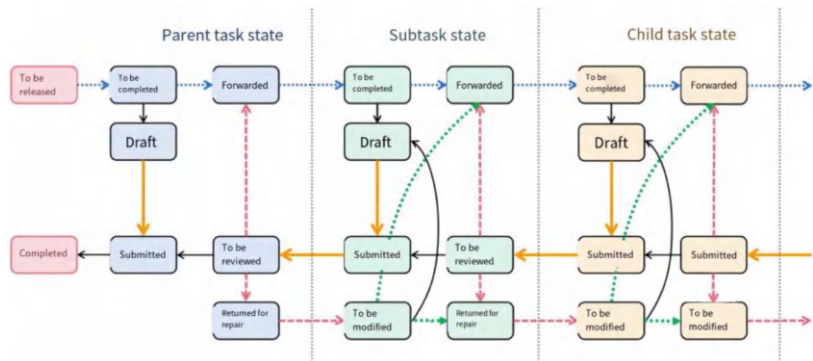


Fig. 1. Task Status Diagram

3.2 Detailed Design of System Core Functions

This article designs a universal data collection system underlying data storage scheme by adopting a knowledge element storage model that separates patterns and data. In terms of specific implementation, the data structure (schema) is stored in the form of fields in a relational table, while the collected data is stored by serializing LOBs. The schema of each data collection task consists of multiple fields, and the detailed information of these fields is saved in the “field table”. A one-to-many relationship is formed between tasks and task detail tables. In addition to storing fixed fields, the task detail table stores all other information through LOB serialization. This design effectively supports dynamic data collection, ensuring the flexibility of data and the universality of the system.

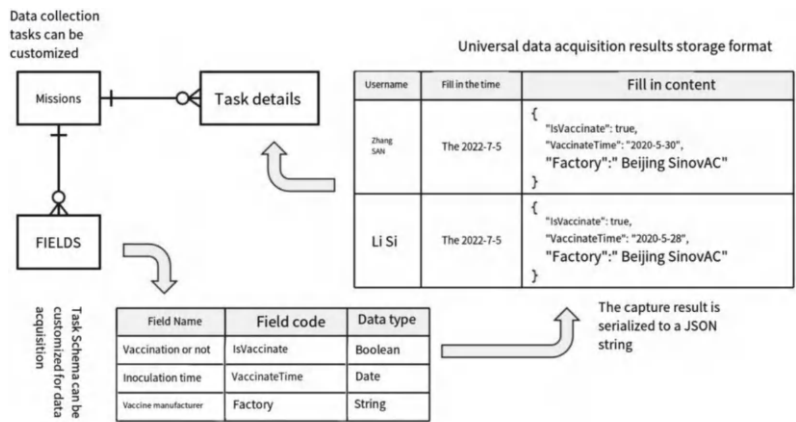


Fig. 2. Design of Non relational Data Relationship Storage Scheme Based on LOB

In order to achieve universality of data storage, this study proposes an innovative data storage solution. In this scheme, the “field value” pairs involved in the data collection task are first converted into “Key Value” format data objects, and then serialized into JSON strings through the “Object Relationship Structure Pattern” in the enterprise application architecture, and stored in the “Task Details” table. When retrieving or displaying data, the stored string can be restored to a data object through deserialization. Modern databases have provided comprehensive support for storing JSON data in NoSQL format in relational databases. For example, the SQL 2016 standard has added query functionality for JSON fields, and SQL Server 2016 and later versions have also introduced JSON functions to support processing JSON data in relational databases, enabling analysis of JSON documents in relational structures or conversion of relational data into JSON text. As shown in Fig. 2.

In general data collection systems, tasks often go through multiple layers of forwarding, summarization, and hierarchical review. In order to effectively support this process, we adopted the design method of self association tables to transform the traditional tree based task decomposition model into a structure suitable for relational storage. We have designed multi-level task forwarding schemes for different data collection modes, especially for form filling and additional classes. In the form filling mode, the task details are initially fixed, and users can only forward tasks and cannot change their content until the task is completed. In this scheme, the forwarding process of tasks is recorded through self association relationships, ensuring clear paths for task level by level forwarding, and controlling data access permissions and editability through task status. Relatively speaking, the additional data collection mode allows task detail records to be initially empty, and the filling person can dynamically add, modify, or delete records, but can only manipulate their own data to prevent others from tampering. This design ensures the accuracy and security of data through strict permission control. During the forwarding process, the forwarded party is granted the permission to add new records, and the status of the task is separately recorded in the task details table to ensure data integrity and traceability, ultimately reviewed by the task initiator. As shown in Fig. 3.

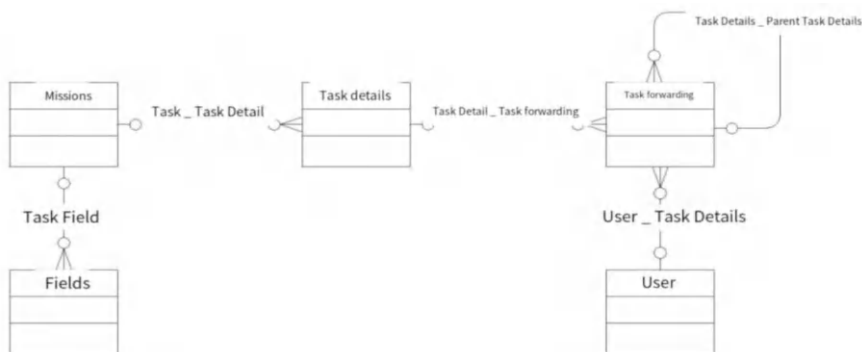


Fig. 3. Logical database design for multi-level forwarding function of data collection tasks

3.3 Database Design

In the design of the main business model of the database, it is necessary to comprehensively consider the multi-level forwarding requirements of table filling and additional data collection tasks. To ensure that the system can support both modes simultaneously, we achieve this goal by merging different conceptual models. A new entity called ‘Task Type’ has been introduced, which determines whether different fields in the task forwarding table are allowed to be empty, thereby distinguishing and managing between different task types. This design helps to better control the task flow process and ensure that each task can be completed smoothly according to the established process.

In database architecture design, the virtual table structure (schema) is the foundation of the design, while meta knowledge is implemented through specific field combinations. By introducing meta knowledge management, different data collection schemes can share the same field combinations, thereby improving the cross integration capability of data. This design allows the system to flexibly handle different types of data collection tasks and better adapt to new requirements when business changes occur. Users can create multiple meta knowledge and architectures in the system to flexibly respond to various business scenarios.

The organizational structure in the system is stored through “organization” entities and forms a tree structure through self association relationships, allowing users to belong to one or more organizations simultaneously. Administrators are responsible for managing the construction of the organization and the allocation of members. In addition, the system also supports user-defined groups and grants them full management permissions, which enables groups to have higher flexibility in specific tasks. This feature not only meets the needs of different business scenarios, but also enables the organizational structure to be quickly adjusted and optimized according to actual situations.

4 Results and Discussion

4.1 Data Access Based on Enterprise Application Architecture

Transaction scripts encapsulate business logic into independent processes, enabling the system to perform various operations through calls to the database. These scripts can not only aggregate multiple database access requests, but also handle different business requirements. To achieve this goal, this system adopts two transaction script implementation methods, based on SQLHelper class and Dataset Query, respectively, to meet different database operation requirements.

Firstly, the SQLHelper class encapsulates common methods in database operations, such as VNet, RunScalar, RunNoQuery, and ExecutStoreProc. It handles different operational scenarios, such as returning data tables, obtaining scalar values, executing non query commands, and executing stored procedures. In this way, transaction scripts can flexibly execute dynamically generated SQL commands. However, this method also carries certain risks, especially when parameterized queries are not used, it is prone to SQL injection attacks. When handling transactions containing multiple SQL commands, additional classes can be written for encapsulation and transaction mechanisms can be used to ensure data consistency and security.

Secondly, the typed dataset provided by ASP.NET allows developers to design transaction scripts through a graphical interface, thereby improving development efficiency. However, due to the inability of typed datasets to handle dynamic SQL commands, complex business scenarios often require the use of SQLHelper classes to jointly encapsulate transaction scripts. Although dataset queries generate SQL commands through a graphical interface, fundamentally it is still an encapsulation of SQL commands. When creating a query, Visual Studio generates a '. Designer. cs' file containing automatically generated SQL command call code. When multiple SQL commands need to be executed or further operations need to be performed after processing SQL results, the results can be transformed into appropriate data types by encapsulating the dataset query methods, such as converting empty integers to boolean types.

```
public class Users
{
    public static bool IsUserInRole(string UserID, string RoleName)
    {
        DAL.DSUsersTableAdapters.UserHelper helper = new
        DAL.DSUsersTableAdapters.UserHelper();
        int? result = helper.IsUserInRole(UserID, RoleName);
        if (result.HasValue && result.Value > 0)
            return true;
        return false;
    }
}
```

To handle the execution of multiple SQL commands, the method of secondary encapsulation can be used. The following is a simplified code example, showing the implementation of the transaction script for assigning roles to users while creating them:

```
public static void CreateUser(string UserId, string LoginName, string Name, string
Password)
{
    DAL.DSUsersTableAdapters.UserHelper helper = new
    DAL.DSUsersTableAdapters.UserHelper();
    if (helper.IsLoginNameExist(LoginName) != 0)
        throw new Exception("The login name already exists, please change to another login
name");
    helper.CreateUser(UserId, LoginName, Name, Password);
}
```


4.2 Implementation of Data Collection Based on Serialized Object Storage

To ensure the universality of the data collection system, a serialized object storage scheme based on JSON format was designed. The key technical challenges faced in implementing this solution include: determining the storage format of JSON data; How to efficiently read, display, edit, and transmit data during the data collection process; And how to securely and efficiently store the returned JSON data.

In order to achieve flexible data collection and storage, this system allows users to customize a data schema that covers fields of multiple data types. Fields are divided into two categories: read-only and editable. The initial value of the read-only field is set by the task publisher and the filling personnel do not have the authority to modify it; And editable fields allow the filling personnel to edit and update during the collection process. Specifically, for additional record type data collection tasks, all fields can be edited, and users can also delete information that has been filled in before the data is approved.

After the data collection is completed, the information filled in will be stored in JSON format in the “Task Details” table, which includes various field data filled in by the user. By converting the data of each field into key value pairs, JSON format can efficiently store data and also facilitate deserialization in subsequent operations. The stored JSON data will be presented to the user through a web client, and the data filled in by the user will interact through dynamic data binding between the front-end and back-end to ensure accurate data transmission and timely updates to the server.

In the implementation process, the Vue framework played an important role by supporting bidirectional binding between data and controls, allowing users to automatically update the values of client controls when modifying data, reducing the complexity of data feedback. In order to transmit data to the server, the system uses the Axios library to send requests, utilizing ASP The Web API feature of. NET allows the client to send serialized JSON data and primary key information to the server. After receiving data, the web API will store it in the database to ensure the integrity and accuracy of the data.

4.3 General Multi-Level Task Forwarding Based on Tree Structure

When designing a multi-level task forwarding scheme based on a tree structure, the system adopted an ID and ParentID self association table structure, successfully achieving multi-level task forwarding and tracking. Meanwhile, similar methods are also applied to the storage of organizational structures, supporting any hierarchical tree structure. However, traditional recursive query methods have the problem of low efficiency when implementing functions such as tracking or viewing task completion personnel during task forwarding. In order to optimize this process, the system has added Level and Path fields and automatically updated them through triggers, effectively solving the problem of inconvenient operation of tree structures in relational databases.

As shown in Table 1, the Level field is used to record the hierarchical relationships in the tree structure. The Level of the root node is 0, and as the level increases, the Level value gradually increases. The Path field is a path composed of the primary key values of each node, separated by a “.” symbol, representing the complete path from the root node to the current node. This structure makes the process of computing task forwarding

more convenient, and users only need to break down the Path field through separators to easily obtain the complete forwarding path and primary key information of the task.

Table 1. Introduces Level and Path fields to simplify the operational complexity of tree structured relationship storage

MissionId	ParentMissionId	Name	Level	Pat
6	Null	A	0	.6
7	6	B	1	.6.7
8	7	C	2	.6.7.8
9	7	D	2	.6.7.9
10	7	E	2	.6.7.10
11	8	F	3	.6.7.8.11

Through triggers, the values of the Level and Path fields can be automatically maintained, thereby avoiding the need for users to manually write additional code to update these two fields. The code for inserting and updating triggers will ensure that the Level and Path fields are properly maintained during insertion or update. The following is a code demonstration of how to maintain these two fields through triggers, using an organizational table as an example.

insert triggers

```

AS
BEGIN
DECLARE @numrows INT
SET @numrows = @@ROWCOUNT
IF @numrows > 1
BEGIN
RAISERROR('Only supports single line insertion operation !',16,1)
ROLLBACK TRAN
END
ELSE
BEGIN
UPDATE E
SET
HierarchyLevel = CASE
WHEN E.ParentBranchId IS NULL THEN 0
ELSE Parent.HierarchyLevel + 1
END,
FullPath = CASE WHEN E.ParentBranchId IS NULL
THEN ''
ELSE Parent.FullPath END
+ Cast(E.BranchId AS VARCHAR(10)) + ''
FROM Branches AS E
INNER JOIN inserted AS I
ON I.BranchId = E.BranchId
LEFT OUTER JOIN Branches AS Parent
ON Parent.BranchId = E.ParentBranchId
END
END
go

```

Update trigger

```

CREATE TRIGGER tg_BranchUpdate ON Branches FOR UPDATE
AS
BEGIN
IF @@ROWCOUNT = 0
RETURN
IF UPDATE(ParentBranchId)
BEGIN
UPDATE E
SET
HierarchyLevel = E.HierarchyLevel - I.HierarchyLevel +

```

```

CASE WHEN I.ParentBranchId IS NULL THEN 0
ELSE Parent.HierarchyLevel + 1 END,
FullPath = Isnull(Parent.FullPath, '')
+ Cast(I.BranchId AS VARCHAR(10)) + ''
+ RIGHT(E.FullPath, Len(E.FullPath)
– Len(I.FullPath))
FROM Branches AS E
INNER JOIN inserted AS I
ON E.FullPath LIKE I.FullPath + '%'
LEFT OUTER JOIN Branches AS Parent
ON I.ParentBranchId = Parent.BranchId
END
END
go

```

5 Conclusion

This article focuses on the research of a general data collection system architecture based on relational database technology, and proposes a flexible and efficient solution to meet the dynamic and diverse data collection needs. By constructing a knowledge element model, a universal underlying storage architecture was designed, and serialized LOB objects were used to address the shortcomings of relational databases in storing unstructured data. At the same time, the use of self associative logic models enables the transmission and tracking of multi-level tasks, effectively meeting the flexibility and hierarchical management requirements of data collection. The system is based on ASP Developed using NET WebForm technology, combined with Vue framework to dynamically generate interfaces, and implemented data transmission and storage through Web API, innovatively addressing the issues of dynamic data collection and interface matching. In addition, the introduction of automatically updated Level and Path fields has optimized the management performance of tree structured data and significantly improved the system's multitasking capabilities. However, the efficiency shortcomings of relational databases in tree and graph data processing still exist. In the future, NoSQL databases can be combined to explore multimodal storage models, organically integrating relational databases with document databases and key value databases to further enhance the efficiency and scalability of the system. This study provides important references for the development of data collection systems and lays a theoretical and practical foundation for subsequent technological optimization.

References

1. Du, J., Zhang, J., Wu, S.: Credit risk prediction of telecom users based on model fusion. In: IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference. IEEE (2021) <https://doi.org/10.1109/IMCEC51613.2021.9482012>
2. De Paulo, M.C.M., Marques, H.A., Feitosa, R.Q., et al.: New encoder-decoder convolutional LSTM neural network architectures for next-day global ionosphere maps forecast. GPS Solut. (2023)

3. Qin, W.: Research on financial risk forecast model of listed companies based on convolutional neural network. *Sci. Programm.* 2022(Pt.6) (2022)
4. Wei, R., Ding, D.: Problems and countermeasures of financial risk in project management based on convolutional neural network. *Comput. Intell. Neurosci.* **2022**, 1978415 (2022). <https://doi.org/10.1155/2022/1978415>
5. Nagarajan, R., Jothi, J.: Analysing traveller ratings for tourist satisfaction and tourist spot recommendation. *Int. J. Bus Intell. Data Min.* **20**, 208–234 (2022). <https://doi.org/10.1504/ijbidm.2022.10034520>
6. Gu, F., Jiang, K., Hu, X., et al.: Deep learning-based image geolocation for travel recommendation via multi-task learning. *J. Circ. Syst. Comput.* (2022). <https://doi.org/10.1142/S0218126622501274>
7. Zhang, T.T., Ye, Y.: Big Data Analytics for Network Level Short-Term Travel Time Prediction with Hierarchical LSTM and Attention (2022). <https://doi.org/10.48550/arXiv.2201.05760>
8. Amegnaglo, C.J., Mensah-Bonsu, A., Asomanin Anaman, K.: Use and economic benefits of indigenous seasonal climate forecasts: evidence from Benin, West Africa. *Clim. Devel.* **14**(10), 909–920 (2022). <https://doi.org/10.1080/17565529.2022.2027740>
9. Krishnan, A., Padmaja, M.: External commercial borrowings and outward foreign direct investment: evidence from Indian manufacturing firms. *Asian Econ. Lett.* (2023). <https://doi.org/10.46557/001c.74858>
10. Ji-Liang, M.A., Fan, L.I., Hui-Jie, Z., et al.: Commercial cash crop production and households' economic welfare: evidence from the pulse farmers in rural China. *J. Agric. Sci.: Engl. Ed.* **21**(11), 3395–3407 (2022)



System Design and Implementation of Particle Filter Algorithm Combined with Mean Shift in High-Precision Event Camera Positioning

Shu Xu, Erlan Wang, and Haiming Zhang^(✉)

College of Electronic Information Engineering, Wuhan Donghu University, Wuhan 430200,
Hubei, China

18163526293@163.com

Abstract. In order to solve the problems of limited dynamic range, high delay and motion blur in indoor high-precision positioning technology, this paper proposes a particle filter (PF) algorithm combined with the mean shift (Mean Shift, MS) strategy. Through the asynchronous data processing characteristics of the event vision sensor, a high-precision positioning and target tracking system is designed. The event vision sensor is based on recording changes in pixel light intensity. Compared with traditional imaging equipment, it has the advantages of low latency, high resolution, and wide dynamic range, and exhibits excellent performance in high-speed motion and complex lighting environments. Through an in-depth analysis of the positioning technologies supported by different types of receivers and their shortcomings, this paper proposes a new positioning method with event data as the core, and designs a solution to address the limitations of traditional positioning methods in terms of delay, interference and accuracy. Optimize the model, thereby significantly improving the accuracy, real-time performance and robustness of positioning. Experimental verification shows that the designed system can maintain centimeter-level positioning errors under a variety of dynamic conditions, and shows strong adaptability in fast target motion and high dynamic range scenarios. This paper combines the mean shift strategy with particle filtering to achieve efficient tracking of target trajectories in complex scenes by optimizing the particle distribution update mechanism. It further verifies the fast convergence and real-time performance of this method under dynamic conditions, providing indoor accuracy. The field of positioning and target tracking provides reliable theoretical basis and practical reference.

Keywords: Event Camera · Indoor Target Tracking · Mean Shift Algorithm · Particle Filter Optimization · Dynamic Data Processing

1 Introduction

In recent years, with the rapid development of wireless communication technology and the widespread popularity of smart terminals and wearable devices, indoor positioning technology has ushered in important technological breakthroughs, which has greatly

promoted the rapid growth of the location-based services (LBS) industry. In application scenarios such as smart homes, smart buildings, automated production lines, and confined space rescues, the demand for high-precision positioning is increasing day by day, and location information has become one of the important core driving forces in the digital economy era. Azure Map (2020) launched by Microsoft provides innovative solutions for indoor augmented reality navigation, while Alibaba's AliBeacon technology has greatly improved the convenience of shopping mall navigation and self-service payment. Since the implementation of the "13th Five-Year Plan", China has regarded navigation and positioning technology as a key research and development direction, and further clarified the strategic goal of improving satellite positioning and navigation capabilities in the "14th Five-Year Plan" to promote the sustainable development of the location-based service industry. Develop.

In the field of indoor positioning technology, multiple solutions have emerged, such as Global Navigation Satellite System (GNSS), Wi-Fi, Bluetooth, Radio Frequency Identification (RFID), and ultrasound. Traditional GNSS technology has low positioning accuracy due to signal obstruction and multipath effects in indoor environments, making it difficult to meet high-precision positioning requirements. The positioning accuracy based on Wi-Fi, Bluetooth and radio frequency technology can usually only reach the meter level. Although there have been improvements, it still cannot meet the needs of some high-precision scenarios. Although ultrasonic positioning can provide decimeter-level accuracy, its high hardware requirements limit its application in some scenarios.

Visible light positioning technology (VLP) is gradually attracting attention as an emerging solution. VLP technology combines the advantages of visible light communication systems, has the characteristics of no radio frequency interference, and provides lighting and communication functions at the same time. It is especially suitable for environments with strict electromagnetic interference requirements, such as hospitals, laboratories and aircraft cabins. However, VLP technology will be affected by changes in lighting conditions in practical applications, and may cause positioning delays and image blur problems in dynamic scenes. With the commercialization of event camera technology, such as the DAVIS series launched by iniVation and the CeleX series of products from OmniVision, it is possible to provide technical support for application scenarios such as visual tasks. Table 1 shows a comparison of relevant parameters of several currently common commercial event cameras.

To address these challenges, this paper proposes a visible light localization method based on event cameras. The event camera captures pixel brightness changes through an asynchronous triggering mechanism. It has ultra-high temporal resolution and wide dynamic range (120 dB), which has significant advantages in high-speed motion and high-dynamic environments. After introducing the particle filter algorithm, the event camera can achieve stable tracking even under occlusion, thereby improving positioning accuracy and real-time performance.

Table 1. Typical event camera technical parameter data

Device model	DVS	ATIS	DAVIS240	DAVIS346	CeleX-V
Maximum frame rate (fps)	-	-	35	40	100
Response Delay (μ s)	15	3	3	20	8
Power Consumption (mW)	23	50–175	5–14	10–170	400
Wide dynamic range (dB)	120	143	120	120	120
Area per pixel (μm^2)	40×40	30×30	18.5×18.5	18.5×18.5	9.8×9.8
Resolution (number of pixels)	128×128	304×240	240×180	346×260	1280×800
Chip area (mm^2)	6.3×6	9.9×8.2	5×5	8×6	14.3×11.6

2 Related Research

In recent years, event cameras, as a neuromorphic vision sensor with high temporal resolution, high dynamic range and low latency, have made significant progress in application research in many fields. In terms of event data denoising, Lin Wanmin [1] proposed a two-step denoising algorithm GMCM, which combines Gaussian preprocessing with motion denoising to greatly improve the signal-to-noise ratio and computational efficiency under high-noise conditions. It was also used on the DVSCLEAN data set Achieved leading denoising effect. In the field of non-line-of-sight imaging, C Wang proposed a passive non-line-of-sight imaging method based on event data [2]. It obtains dynamic information through asynchronous event streams, significantly alleviates the degradation problem caused by moving targets, and builds the first event non-line-of-sight imaging method. Imaging data set EM-NLOS, experiments have verified that this method significantly improves the signal-to-noise ratio and perceptual similarity.

In eye tracking research, P Bonazzi [3] developed a neuromorphic eye tracking method based on dynamic vision sensor (DVS) event data, through the combination of a lightweight spiking neural network model “Retina” and a low-power edge processor Speck, achieving high-precision pupil positioning with an error of only 3.24 pixels, power consumption as low as 2.89–4.8 mW, and delay only 5.57–8.01 ms, showing strong energy efficiency advantages. Z Liu proposed a calibration method based on geometric lines in event camera calibration [4]. It detects lines directly from the event stream and combines it with a nonlinear optimization algorithm. It can efficiently complete monocular and dual-object calibration without the need for additional calibration objects and has strong adaptability. And has a wide range of applications.

In the field of event camera simulation, D Joubert designed an improved DVS pixel simulator [5], which is closer to the actual sensor behavior by simplifying the delay and noise model and adding readout circuit modeling, and verified its performance in sensing speed on the dynamic MNIST data set and energy saving potential. J Kim developed

the N-ImageNet large-scale data set, which simulates event data collection through programmed hardware [6], covering different camera trajectories and extreme lighting conditions, providing a challenging benchmark for fine-grained object recognition by event cameras. Experiments show that pre-training based on this data set significantly improves the performance of the event classifier, especially the learning ability under the condition of a small amount of labeled data.

In terms of network structure innovation, M Gehrig proposed Recursive Visual Transformers (RVTs) [7], which introduced convolution priors, local and dilated global self-attention, and recursive temporal feature aggregation through multi-stage design, and implemented it on the Gen1 automobile data set. It achieves 47.2% mAP, while the inference time is as low as 12ms, and the parameter efficiency is increased by 5 times, providing an efficient solution for event camera object detection. In addition, RW Baldwin proposed the TORE event representation method [8], which significantly optimizes memory and computing efficiency in noise reduction, image reconstruction and classification tasks by compactly storing event time information, avoiding time blocking and data expansion.

In the research on self-supervised learning of event cameras, F Paredes-Vallés used an unsupervised intensity reconstruction method for the first time [9], combining optical flow estimation with event-based photometric consistency without relying on real or synthetic data, and proposed an efficient and lightweight optical flow estimation. The network is close to the current optimal performance while maintaining a high inference speed. M Gehrig further introduced feature correlation and sequence processing in optical flow estimation, significantly improving the accuracy of dense optical flow calculations [10]. Compared with existing methods, the endpoint error was reduced by 23%, and it built a model that includes larger displacements and higher A new data set with high resolution, on which the endpoint error is reduced by 66%.

In driver monitoring systems (DMS), C Ryan proposed a new method to simultaneously detect and track the driver's face and eyes using the high temporal resolution of event cameras [11]. Through a fully convolutional recurrent neural network and Neuromorphic-HELEN synthetic event data set, this method achieves precise eye movement analysis, especially showing unique advantages in driver fatigue detection based on eyelid flicker behavior.

Event camera technology has made extensive and in-depth research progress in the fields of data denoising, imaging, classification, optical flow estimation, calibration and driver monitoring. These studies provide important theoretical support and technical reference for this paper to further optimize the application of event cameras.

3 Data-Driven Positioning Method Based on Event Camera

3.1 Characteristics and Signal Processing Methods of Event Camera Data

Event cameras differ from traditional cameras in that they output event stream data by capturing dynamic changes in the scene in real time, with each event including timestamp, pixel position and polarity. This method has higher temporal resolution than traditional images and is suitable for high-speed and low-light environments. However, event stream data has noise problems in practical applications. These noises are different

from the noise mechanisms of traditional images, and mainly manifest as background noise, noise caused by threshold fluctuations, and hot spot noise.

Background noise refers to pixels inside the camera that still trigger events without an external event source. This type of noise is mainly caused by factors such as circuit instability, such as thermal noise and charge injection effects. In experiments, it was found that such noise often occurs around stationary LED light sources, and its distribution exhibits Poisson characteristics. The noise generation rate is generally between 0.03 and 0.2 events/pixel/second, and will be affected by the environment and equipment.

Threshold fluctuations are also the main cause of noise. The event camera triggers events based on the threshold of light intensity changes. However, due to threshold fluctuations, events may not be triggered in time or may be triggered early before the threshold is reached. Threshold fluctuations usually appear as normal distribution, and their fluctuation amplitude is generally between 2% and 4% of the threshold. This instability affects the accuracy of event stream data to a certain extent. Hotspot noise is similar to dead pixels in traditional cameras, where certain pixels continue to generate events due to hardware failures or circuit issues. These abnormal events will affect data quality and may interfere with the normal operation of the system in severe cases. Hotspot noise is usually caused by damaged pixels or electrical interference.

Based on the previous analysis of the noise characteristics and distribution of event cameras, this study selected the frequency range from 600 to 1500Hz in LED beacon detection, aiming to reduce background interference caused by camera movement and improve the accuracy of frequency detection. On this basis, an LED detection method based on beacon frequency mapping is proposed. This method uses the LED-ID frequency data collected by the event camera as input, and finally outputs the pixel coordinates corresponding to the frequency of the LED beacon. In the entire processing process, polarity conversion events are first generated, then the time intervals are calculated, and finally a frequency map is generated based on the time intervals. Through these steps, the precise positioning of the LED beacon is achieved.

Each data point of the event sequence includes timestamp, pixel coordinates and polarity information, recording the timing of light intensity changes. When the polarity of the current event is different from the polarity of the previous event, a polarity conversion event is generated, and the timestamp and pixel coordinates of the event are recorded. Through this process, the raw event sequence is converted into a set of polarity switching events. This transformation helps reduce background noise and provides useful time series data for subsequent frequency calculations.

Next, based on the polarity transition event, the time interval between adjacent events is calculated. For each pixel, the time difference between two adjacent polarity switching events is used to estimate the beacon's blink frequency. Assuming that the time interval is Δt , where t_i represents the timestamp of the i event, the time interval between two events can be expressed as formula (1).

$$\Delta t_{i,i+1} = t_{i+1} - t_i \quad (1)$$

The calculated time interval data provides a preliminary basis for frequency estimation, but due to the influence of noise, the accuracy of frequency estimation is still limited. In order to further improve the estimation accuracy, this study proposes a frequency map generation method based on interval data. By calculating the reciprocal

of each time interval, the frequency distribution of the LED beacon can be obtained. Assuming f is the frequency, the frequency can be estimated from formula (2) through the relationship between the reciprocal interval Δt and its probability density function $p(\Delta t)$.

$$f = \frac{1}{\Delta t}, p(f) = p(\Delta t) \quad (2)$$

An accurate frequency map is generated by taking a weighted average of the frequency values across all event intervals and adjusting the weights based on the assumption of a Gaussian distribution. The generation of this map not only relies on the frequency distribution of each time interval, but also combines the detailed analysis of the frequency characteristics in the experimental results, thereby achieving the precise positioning of the LED beacon on the pixel coordinates.

3.2 Positioning Algorithm and Error Analysis Based on Event Data

In the design and implementation of a high-precision event camera positioning system, this method combines the information of multiple LED lamps and estimates the positioning based on event data. We set the world coordinates of the LED lamps and assume that the imaging plane is parallel to the ceiling, ignoring the rotation effect between the camera coordinate system and the world coordinate system, thereby simplifying the positioning calculation process. With the known center position and focal length of the event camera lens, we can deduce the physical distance between the LED lamps and then the position of the camera. Combining the relationship between the image coordinates and the actual world coordinates, the system calculates the position of the LED lamp on the image plane to obtain the required data for precise positioning.

Although particle filtering performs well in improving positioning accuracy, it is still affected by distortion errors. Camera distortion usually originates from defects in lens shape or asymmetry of installation angles. Radial distortion and tangential distortion are the two most common types of distortion. In order to reduce the impact of distortion on positioning accuracy, this paper introduces a camera distortion model and compensates it through calibration. Radial distortion mainly occurs at the edge of the image, caused by the difference in magnification of different parts of the lens; while tangential distortion is caused by the non-parallelism between the imaging plane and the lens installation plane. In practical applications, the distortion coefficient is obtained through precise camera calibration, and the camera is fully calibrated using Zhang Zhengyou's chessboard method to ensure the high accuracy of the calibration results. Finally, based on the calibration results shown in Table 2, the positioning algorithm is optimized, the error caused by distortion is effectively reduced, and the accuracy of camera positioning is further improved.

The system installs four LED light sources at different positions on the ceiling and uses the pulse signal of each LED for positioning. The signal frequency of LED is between 500 Hz and 1000 Hz, the duty cycle is 50%, and the positions are LED-1 (200, 100, 1000), LED-2 (500, 100, 1000), LED-3 (200, 500, 1000), LED-4 (500, 500, 1000), the distance between the camera lens and LED is 1000 mm, and the positioning area is 700 mm \times 700 mm. First, the intrinsic parameters and distortion parameters of the

Table 2. Calibration result data

Parameter	Calibration results
Image resolution	346 pix \times 260 pix
Tangential distortion coefficients p_1, p_2	0.0010534, -0.10838
Focal length/pixel size	272.121, 271.331
Radial distortion coefficients k_1, k_2, k_3	$-0.44701, 0.2756, -0.0009001$
Cell size	18.5 $\mu\text{m} \times 18.5 \mu\text{m}$
Principal coordinates (u_0, v_0)	(179.423, 138.704)

camera are obtained using the chessboard calibration method, and the focal length is 12 mm. By analyzing the signal frequency distribution of each LED, the maximum frequency value is calculated, and the world coordinates of the LED are matched with the pixel coordinates by geometric transformation, so as to estimate the camera position.

In order to optimize the positioning results, this paper introduces a median selection strategy, which calculates the median by combining multiple LEDs to reduce error fluctuations, and finally stabilizes the positioning error within 2 cm. 15 test points were selected in the experiment to verify the effectiveness and stability of the method. The positioning accuracy after optimization is significantly improved, the errors caused by environmental interference and data anomalies are reduced, and a high positioning accuracy is ensured. The positioning method combined with the mean shift particle filter algorithm shows good accuracy and robustness in the environment of multiple LED signal sources, and has wide application potential.

4 Application and Optimization of Particle Filter Algorithm in Dynamic Positioning

4.1 Principle of Particle Filter Algorithm and Positioning Accuracy Improvement

In the precise positioning process of event cameras, particle filter (PF) as a nonlinear estimation algorithm has been widely used in the estimation of dynamic system states. This method is particularly suitable for tracking and positioning each LED signal in the image coordinate system. The basic idea of particle filtering is to generate a group of particles near the position to be tracked, and use the event data obtained by the sensor to update the state of the particles, so as to obtain the optimal estimate of the target position.

Specifically, in the LED positioning scene, the particle filter first randomly distributes the particles in the image coordinates. Then, the weight and position state of the particles are adjusted by calculating the beacon frequency mapping relationship between each particle and the event data. This process estimates the optimal position of the target through weighted averaging, and combines the information of the world coordinate system to finally achieve accurate positioning and real-time tracking of the event camera.

In order to verify the application effect of particle filtering in nonlinear dynamic systems, this study designed a typical one-dimensional system model for testing. The state

and observation equations of the model are shown in formulas (3) and (4) respectively.

$$x_k = x_{k-1} + 0.5 \cdot \cos(1.2 \cdot k) + v_k \tag{3}$$

$$z_k = x_k + v_k \tag{4}$$

wherein, v_k and v_k are Gaussian noises with a mean of zero and a variance of 1. In order to verify the effectiveness of the algorithm, the experimental results show that the positioning accuracy of the combination of particle filtering and mean shift algorithm is significantly better than the traditional particle filtering method, especially in a dynamically changing complex environment, the system can effectively reduce the positioning error and enhance the robustness. The performance of the DAVIS346 event camera used in the experiment under different focal length and field of view configurations shows that, as shown in Table 3 and Figs. 1 and 2, the focal length and field of view have an important influence on the positioning accuracy, and the appropriate viewing angle configuration can significantly improve the performance of the positioning system.

Table 3. Focal length and field of view parameters of event camera DAVIS346

Focal length (L) [mm]	Horizontal angle AFOV [deg]	Vertical view AFOV [deg]
2.1	113	97.7
3.5	4.5	70.8
4.5	70.8	56.2
6	56.2	43.7
12	29.9	22.7

By combining particle filtering with the mean shift algorithm, not only the accuracy of the positioning system is improved, but also its adaptability in dynamic environments is enhanced. This method provides a new high-precision positioning technology path for autonomous driving, robot navigation and other fields, and has broad application prospects.

4.2 Multi-Source Data Fusion and Error Correction Based on Particle Filtering

In the particle filter (PF) algorithm, the target tracking task is first started by initializing the particle swarm. The particle swarm is evenly distributed around the target according to the target’s initial state distribution $P(x_0)$. Usually, 1000 particles are selected to simulate the target’s state. The weight of each particle at the initial moment is equal, which ensures a relatively accurate representation of the target’s true state. As the tracking process proceeds, each particle is updated according to the target’s motion equation, and the weight is adjusted in combination with the observed data $Z(k)$. To achieve this, the frequency maximum of the LED-ID signal is used as the target’s observed position, the

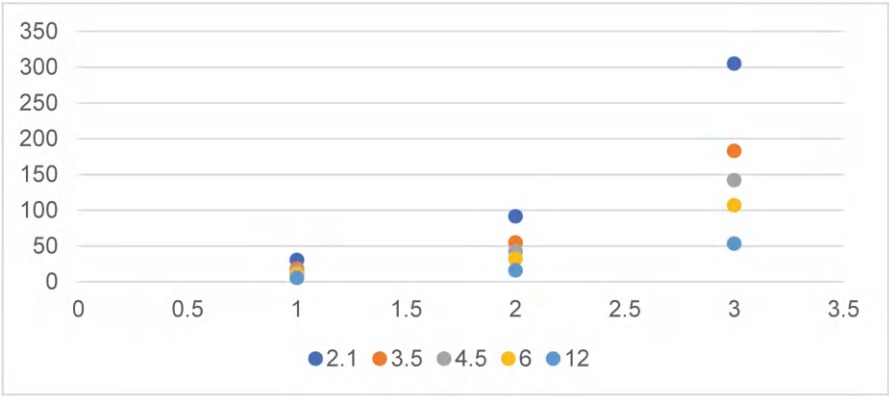


Fig. 1. Horizontal linear field of view and imaging plane parameters of event camera DAVIS346

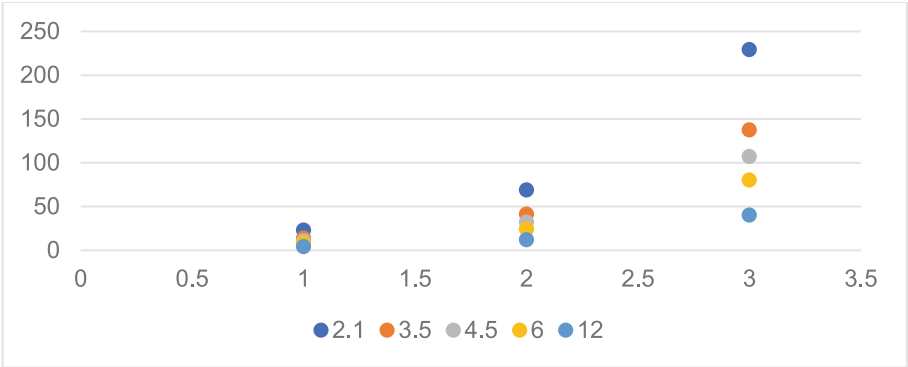


Fig. 2. Event camera DAVIS346 Vertical linear field of view and imaging plane parameters

difference between the particle and the actual observation value is calculated, and then the weight is adjusted by the Gaussian function. In this way, particles that are closer to the target state will receive a higher weight, while particles that deviate from the target will receive a lower weight. Next, the system will resample, retaining those particles with larger weights and discarding those with smaller weights to improve the accuracy of state estimation.

To avoid particle degradation, the weights are normalized after each tracking to ensure that the particle swarm can continue to provide effective state estimation. Based on the normalized weights, the resampling operation will increase the probability of selecting particles with larger weights, thereby improving the estimation accuracy of the target state. The simulation results show that when the initial state of the particle swarm is close to the true state of the target, the tracking error will be smaller; if the initial speed is more accurate, the tracking error will gradually decrease and eventually converge. Through simulation verification under different initial conditions, it can be concluded that the initial state and speed play a key role in improving the accuracy.

The mean shift algorithm is an effective non-parametric clustering method. Its core idea is to drive particles to gather in areas with higher density according to the offset mean of the particles. Under the particle filter framework, the MS-PF algorithm corrects the position of the particles by combining the mean shift correction step before resampling the particles to ensure that their state is closer to the actual distribution of the target. Through repeated mean shift iterations, the particle set gradually approaches the optimal position of the target, thereby improving the estimation accuracy of the particle state. This strategy significantly improves the distribution of particles, avoids the particle degradation problem in the traditional particle filter algorithm, and improves the robustness of the system.

The experimental results show that the MS-PF algorithm can effectively correct the particle state by introducing the Gaussian kernel function as the core of the mean shift. During the experiment, the choice of the number of particles has an important impact on the positioning accuracy and computational efficiency. Specifically, when the number of particles is 1000, the tracking error of the MS-PF algorithm is reduced compared with the traditional particle filter algorithm, and the calculation time is also increased. However, after adding the mean shift algorithm, although the amount of calculation has increased, the calculation efficiency is improved while ensuring the accuracy, especially when the number of particles is small, the real-time performance of the MS-PF algorithm has been significantly improved.

Further analysis shows that when the number of particles is reduced, the performance of the MS-PF algorithm is more advantageous than the traditional PF algorithm. When the number of particles is 500, the positioning error of the MS-PF algorithm changes little, while the traditional PF algorithm has a large error increase. In addition, the amount of calculation of the MS-PF algorithm under this number of particles is greatly reduced, and the running time is short, indicating that its application in real-time positioning tasks has significant advantages. Table 4 is a detailed comparison of the experimental results, showing the tracking error and total computation time of the PF and MS-PF algorithms under different particle numbers.

Table 4. Tracking error and total computation time of the PF and MS-PF algorithms under different particle numbers

	Particle Filter (PF)	Mean-Drift Particle Filter (MS-PF)	Particle Filter (PF)	Mean-Drift Particle Filter (MS-PF)
Number of particles	1000	1000	500	500
Total Computation Time (s)	1.357	1.796	0.665	0.913
Maximum Tracking Error (cm)	2.0	1.8	2.5	2.2

It can be seen from the above experimental results that the MS-PF algorithm can effectively reduce the tracking error under the same number of particles, and maintain a small error growth when the number of particles decreases. In addition, although the addition of the mean shift algorithm increases the calculation time compared to the traditional PF algorithm, the improvement in its accuracy makes up for the increase in calculation time, especially when the number of particles is small, the calculation efficiency of the MS-PF algorithm is significantly improved, showing good real-time performance and positioning accuracy.

5 Conclusion and Prospect

This study proposes a visible light image positioning method based on event camera, aiming to overcome the problem of insufficient positioning accuracy of traditional CMOS sensors in high dynamic range and high-speed motion scenes. By introducing the event camera and taking advantage of its low latency, high dynamic range and efficient data processing, we can achieve accurate positioning based on event data, thereby significantly improving the positioning accuracy and mobile terminal trajectory prediction capabilities in dynamic environments. In response to the problems of noise and background motion interference in the detection process of LED beacons, we proposed a detection method based on beacon frequency mapping, which can effectively improve the accurate detection of LED beacon pixel positions. Furthermore, we improved the particle filter algorithm and proposed the MS-PF algorithm, which can reduce the computational complexity while improving positioning stability, accuracy and real-time performance. However, despite the progress made in this study, there is still room for improvement in the application of three-dimensional scenes, real-time tracking of targets that are occluded for a long time, and positioning effects in complex outdoor environments. In the future, we will focus on optimizing algorithms, integrating complete positioning systems, and conducting cross-scenario verification to enhance the application potential of this technology in fields such as intelligent navigation and autonomous driving.

Acknowledgements. This work was supported by Research on indoor navigation system base on particle filter algorithm 2020 youth fund of project of Wuhan Donghu University. Youth Foudation WuHan Donghu University 2020dhzk003.

References

1. Lin, W., Li, Y., Xu, C., et al.: A motion denoising algorithm with Gaussian self-adjusting threshold for event camera. *Visual Comput.* (2024). <https://doi.org/10.1007/s00371-023-03183-4>
2. Wang, C., He, Y., Wang, X., et al.: Passive non-line-of-sight imaging for moving targets with an event camera. *Chin. Opt. Lett.* **21**(6), 061103 (2023). <https://doi.org/10.3788/COL202321.061103>
3. Bonazzi, P., Bian, S., Lippolis, G., et al.: Retina: low-power eye tracking with event camera and spiking hardware. *IEEE* (2023). <https://doi.org/10.1109/CVPRW63382.2024.00577>

4. Liu, Z., Guan, B., Shang, Y., et al.: LECalib: Line-based event camera calibration. *Measurement* **235**, 114900 (2024)
5. Joubert, D., Marcireau, A., Ralph, N., et al.: Event camera simulator improvements via characterized parameters. *Front. Neurosci.* **15**, 702765 (2021). <https://doi.org/10.3389/fnins.2021.702765>
6. Kim, J., Bae, J., Park, G., et al.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2146–2156 (2021)
7. Gehrig, M., Scaramuzza, D.: Recurrent vision transformers for object detection with event cameras. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13884–13893 (2023)
8. Baldwin, R.W., Liu, R., Almatrafi, M., et al.: Time-ordered recent event (tore) volumes for event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(2), 2519–2532 (2022)
9. Paredes-Vallés, F., De Croon, G.C.H.E. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3446–3455 (2021)
10. Gehrig, M., Millhäusler, M., Gehrig, D., et al.: E-raft: Dense optical flow from event cameras. In: *2021 International Conference on 3D Vision (3DV)*, pp. 197–206. IEEE (2021)
11. Ryan, C., O’Sullivan, B., Elrasad, A., et al.: Real-time face & eye tracking and blink detection using event cameras. *Neural. Netw.* **141**, 87–97 (2021)



Research on Optimization of Visual Space Fractal Design Algorithm Based on Fractal Geometry and Complex Network Theory

Huimin Chen¹✉, Zhenting Li², and Junlin Zhou³

¹ College of Architecture and Environment, Sichuan University, Chengdu 610225, Sichuan, China

chm_ucb@163.com

² College of Computer Science, Sichuan University, Chengdu 610225, Sichuan, China

³ School of Cyber Science and Engineering, Chengdu 610225, Sichuan, China

Abstract. This article is based on fractal geometry and complex network theory, dedicated to exploring the optimization application of fractal design algorithms in visual spatial design. Through in-depth analysis of fractal theory and sorting out specific design cases, this article proposes two improved design methods based on mathematical models and programming languages, and generates images with unique visual effects in experiments. This article further explores the expressive power of fractal structures in visual presentation and their multi-level aesthetic connotations, with a particular focus on their unique characteristics in the combination of ordered beauty and psychedelic beauty. This study not only expands the application boundaries of fractal algorithms in visual spatial design, but also provides innovative theoretical support for design methods.

Keywords: Fractal Geometry · Visual Spatial Design · Complex Network Theory · Optimization of Fractal Design Algorithms · Image Generation

1 Introduction

Fractal geometry is a mathematical theory with wide application potential, which reveals the self-similarity of nature and attracts much attention. It has played an important role in both science and art since it was proposed. People can describe the similarity patterns of things in different scales through fractal geometry to help people understand natural phenomena, and also provide theoretical support for structural modeling in many disciplines. Complex network theory provides a way to describe multiple relationships in a system so that the structure and properties of complex systems can be quantified and visualized. The combination of fractal geometry and complex network theory in visual design can expand the expression of design, and provide systematic generation and optimization strategies for design to achieve more hierarchical and harmonious visual effects.

The self-similar structures generated by fractal geometry in visual design applications are favored for their order and aesthetic properties, and complex network theory

also shows unique advantages in strengthening the logical relations of design. Through the self-recursion characteristic of fractal algorithm, designers can create accurate repetitive patterns beyond the traditional hand painting to achieve multi-dimensional visual hierarchy effect. These patterns combined with complex network theory can further optimize the relationship between their elements so that visual works can convey richer information while maintaining the structure. This pattern design combines natural beauty and systematicity visually, and can meet the complex aesthetic needs of modern design in a unique form.

This study focuses on constructing an optimized fractal graphic generation and color control algorithm to improve the innovation and operational convenience of visual space design. By controlling the structure of fractal graphics, designers can generate self similar patterns that meet design requirements without programming, simplifying the traditional manual drawing process; At the same time, the introduction of color control algorithms has given graphics a more flexible way of color presentation, making graphic representation not only hierarchical, but also dynamically adjustable to meet diverse design scenarios. Research on optimizing fractal and complex network algorithms to endow design systems with higher adaptability and controllability, thereby promoting the development of visual design in the digital trend.

The purpose of this article is to inject new expressive methods into visual design by combining mathematics and design theory, helping designers achieve works with mathematical beauty, and providing theoretical basis for the increasingly developed intelligent design. I hope that through this exploration, visual design can not only enrich its expression forms, but also gain a wider application space in the development of digital and intelligent technologies, exploring more possibilities for the deep integration of design and mathematics in the future.

2 Related Research

As a mathematical tool for describing space filling, fractal geometry exhibits uniform geometric properties at different scales through self similarity and recursive fractal patterns. The local details of fractal shapes exhibit similarity when zoomed in, making them an important tool for revealing natural laws, optimizing design structures, and enhancing visual and sensory experiences. Although fractal shapes can be infinitely recursive, in the fields of architecture and design, this theory is often limited by practical scales. J Vaughan explored the application of fractal geometry in architecture [1], emphasizing that fractal forms not only provide support for mathematical aesthetics, but also help designers solve functional problems in architectural space, such as spatial layout, structural optimization, and environmental adaptability. This design thinking that combines natural self similarity characteristics helps create more harmonious and human friendly architectural spaces.

In the generation and processing of art and design images, traditional iterative hard thresholding methods often produce significant artificial effects at low sampling rates, which not only affect visual effects but may also interfere with subsequent image processing. To address this issue, C Shi proposed a one-dimensional MFDMA algorithm [2], which effectively removes most of the artificial effects by extracting features at different

scales and fusing multi-scale information, significantly improving the subjective visual effect of art and design images. This fractal based algorithm not only improves the quality of image restoration, but also makes the structure of the image closer to the original design, providing important technical support for image generation and optimization in art and architectural design.

The widespread application of fractal patterns is not limited to images and art design, and its potential in architecture and urban planning is also constantly being explored. Z Yu simulated different fractal patterns and conducted geometric evaluations to select the most suitable fractal pattern for architectural design [3]. This method not only demonstrates high aesthetic value in visual effects, but also provides architects with more creative and functional design solutions.

However, despite significant progress in fractal computing methods in the fields of architecture and design, the application of fractals in urban planning still lags behind the development of technology. F Jahanmiri pointed out through a review of existing literature and bibliometric analysis that there is still a research gap regarding the direct correlation between planning norms and fractal patterns [4], and most of the relevant literature has been published outside the field of planning.

Fractal is not only a visual aesthetic element, but its design benefits are also reflected in the improvement of psychological and sensory experience. JH Lee explored the relationship between fractal patterns and visual perception by combining fractal analysis and visual attention simulation. He found a correlation between fractal dimensions and pre attention processing of visual stimuli [5], providing a quantitative analysis method for the relationship between visual stimuli and perception in architectural design.

In addition, the application of fractal patterns can significantly improve human psychological and emotional experiences, especially in interior design and the creation of urban environments. Robles K E's research shows that fractal design can not only enhance people's aesthetic identification with the environment [6], but also effectively reduce pressure in space, enhance residents' sense of relaxation and participation.

More in-depth cross sensory research also indicates that the aesthetic effects of fractal design are not limited to the visual level, but can be extended to sensory experience areas such as touch and hearing. C Viengkham's research shows that in multiple sensory domains such as vision, hearing, and touch, people generally prefer $1/f$ structures that are similar to natural scenes [7], which reflect the complexity and self similarity in nature.

With the continuous development of fractal computing technology, AL Schor [8] proposed a new computational method for efficiently generating Mandelbrot type fractals that approximate user-defined shapes. This method not only improves the speed and stability of fractal calculations, but also introduces a new shape modulus function, allowing fractal design to flexibly control shape and accurately adjust fractal details.

AA Briellmann [9] summarized the application of fractal patterns in urban design, emphasizing their positive role in enhancing environmental attractiveness, improving urban walkability, and intuitive navigation. He also proposed a solution that combines fractal elements with biological affinity and traditional architecture, which can effectively promote the health of urban residents, reduce stress, and alleviate psychological fatigue.

J. Friedenberg's experimental research revealed the influence of reflection, rotation, and displacement in fractal patterns on aesthetics, indicating that rotated and reflected patterns are usually more preferred, and as local symmetry decreases, aesthetics also decrease [10]. With the continuous advancement of fractal computing methods, the potential of fractal design will be more widely explored, especially in applications such as cross sensory aesthetics, spatial optimization, and mental health, which will provide important references for future design theory and practice.

3 Research on the Application of Fractal Design Algorithms in Visual Space

3.1 Application and Development of Fractal Design Algorithm

The relationship between mathematics and art has been gradually explored in depth since the Renaissance, especially in the application of perspective. Mathematics provides a scientific basis for artistic creation and promotes the innovation of artistic expression. By mastering mathematical principles, artists can reproduce three-dimensional space more accurately, thereby enhancing the realism of their works. In the 20th century, art schools such as Cubism and Futurism introduced geometry and non-Euclidean geometry into their creations, breaking through the expression framework of traditional art and creating a new artistic perspective. In this period, mathematics not only provided artists with new means of expression, but also promoted artists to explore the deep connection between higher dimensional space and form formation in their works. Advances in mathematics were intertwined with innovations in the arts, further blurring the lines between the two.

In the 21st century, the fractal theory has further deepened the integration of mathematics and art. Fractals provide a unified framework for the interpretation of complex shapes in nature and provide new tools for artists to create more visually impressive works. The development of computer technology has made an unprecedented breakthrough in fractal art. Artists use computer-generated fractal images to combine mathematics with artistic forms to create more complex and rich artistic effects. Fractal theory has moved from an abstract mathematical concept to practical application and has become an important tool in science, art and other fields, further promoting the deep communication and resonance between mathematics and art.

Fractal design algorithm is a method to generate complex graphs through mathematical formulas and computer iterative processing. It combines fractal theory with modern computing technology to create 3D structures with self-similar features. Fractal design was initially applied to the iteration of simple geometric figures, and gradually developed into more complex systems over time. The Cantor triplex forms a self-similar discrete point set by dividing and removing line segments, which shows the infinite complexity of fractal patterns. The Koch curve and snowflake morphology show the complex structure of natural snowflakes by repeatedly dividing the sides of triangles. Sierpinski's base blanket and sponge also further enrich the application of fractal design through the iterative treatment of squares and cubes, respectively, reflecting the wide range of possibilities of fractal structures in two-dimensional and three-dimensional space.

With the progress of technology, fractal design algorithm has been extended to color design to form "color separation design algorithm". This method applies fractal theory to

color changes, and uses computer algorithms to perform regular color iterations to create rich, dreamy color effects. This design approach avoids the monotony and rigidity of traditional color design, making the colors more dynamic and layered, while also showing the unique complexity and infinity of fractals. The color separation design algorithm not only enhances the expressiveness of visual effects, but also gives the design works a unique artistic charm and visual impact, becoming an important innovative tool in modern design.

3.2 Innovative Application of Color Separation Design Algorithm

The fractal design algorithm combines fractal geometry principles with computer programs to generate complex and uniquely structured design patterns through repeated mathematical operations. Unlike traditional design methods, fractal design algorithms iterate on graphics to visually present self similarity and infinite details, effectively simulating complex morphological features in nature. With the continuous development of computer technology, the application scope of fractal design algorithms has gradually expanded, evolving from simple geometric shapes to displaying extremely complex and layered patterns, becoming an important innovative tool in the field of modern design.

The historical development of fractal design algorithms has gone through a gradual evolution process, where the initial fractal shapes were only generated through simple iterations of the basic geometric shapes. For example, the Cantor set forms a pattern with self similarity by uniformly segmenting line segments and removing intermediate parts. This process is very simple but can be infinitely repeated, thus exhibiting the basic characteristics of fractal shapes. The generation of the Koch curve is even more complex, as it generates a snowflake like shape through continuous segmentation and iteration of the triangle edges. This fractal structure is very similar to the snowflake shape in nature in both form and generation process. The two important models proposed by Sierpinski, Sierpinski carpet and Sierpinski sponge, respectively generate two fractal patterns with deep spatial sense through iterations of squares, planes, and cubes. These early fractal models not only contributed to mathematical theory, but also provided highly creative inspiration for modern design.

Nowadays, with the continuous advancement of computer hardware and algorithms, fractal design algorithms can not only generate simple geometric patterns, but also display extremely complex and exquisite visual effects. Designers can easily create intricate details that traditional design methods cannot achieve with the computational power of computers, while maintaining the uniqueness and innovation of their designs. In the current design field, fractal design algorithms are widely used in various fields such as visual arts, architecture, animation, virtual reality, etc. Through continuous iteration and optimization of algorithms, unprecedented complexity and refinement can be achieved in design, promoting innovation and development of various design works.

4 Analysis of Visual Effects and Aesthetic Features of Algorithm Design

4.1 Aesthetics of Mathematical Logic in Algorithm Design

In contemporary visual space design, the combination of fractal geometry and complex network theory not only promotes innovation in fractal design algorithms, but also provides new forms of expression for mathematical logic aesthetics. Fractal geometry generates complex self similar structures through recursive iteration, while complex network theory reveals the influence of local rules on global structures through the relationships between nodes. The combination of these two provides a new creative method for design, organically integrating mathematical rigor with aesthetic expression, and presenting rich levels and changes in visual effects of the work.

The core of fractal design algorithm lies in the application of iterative formulas. The basic iterative formula of the Mandelbrot set is shown in eq. (1).

$$Z_{n+1} = Z_n^2 + C \quad (1)$$

By adjusting the initial value Z_0 and constant C , a graphical structure with self similarity can be generated. Each iteration makes the graphics more refined, creating a unique visual effect. Designers can adjust the number of iterations and constants according to specific needs, generate fractal shapes of different styles, and achieve diverse visual expressions. Another classic fractal structure is the Rulia set, as shown in the iterative formula (2).

$$Z_{n+1} = Z_n^2 + C \quad (2)$$

Unlike the Mandelbrot set, the shapes of the Rulia set are extremely sensitive to the constant C , and with slight adjustments, they can generate completely different forms, which brings more creative space to visual design.

The fractal design algorithm not only relies on precise calculations of mathematical formulas, but is also closely related to complex network theory. Designers can influence the global shape of fractal structures by adjusting the connection patterns between nodes, especially in the design of three-dimensional fractal graphics. Complex network theory helps optimize fractal structures at different levels, making the design more diverse.

Fractal design algorithm has significant advantages over traditional design methods. In traditional design, designers usually express natural objects by simplifying geometric figures (such as circles and squares), but this abstract method lacks in accuracy and natural restoration. The fractal design algorithm can more truly restore the complex shape of nature through fine parameter control and avoid the limitation of oversimplification.

The infinite variation of fractal algorithm provides more possibilities for design. Since the generation of fractal graphics depends on mathematical iteration, slightly adjusting parameters can produce completely different effects, greatly improving the flexibility and creative space of the design. This feature improves design efficiency and enriches visual effects.

The combination of fractal geometry and complex network theory provides a solid mathematical foundation for fractal design algorithms and promotes the application of

mathematical logic aesthetics in visual space design. Designers can achieve efficient creation and rich performance through precise mathematical control and flexible parameter adjustment. The infinite variability of fractal design algorithms has brought new possibilities to the design field and promoted the interdisciplinary integration of mathematics and art.

The combination of mathematical logic and aesthetics plays a crucial role in visual design based on fractal geometry and complex network theory. The core feature of fractal geometry is that its self-similarity generates regular but changeable graph structure through recursive iteration. Each fractal figure is born in the process of rigorous mathematical calculation to show the beauty of symmetry in mathematics, but also through visual effects beyond the monotonous and ordinary to present the unique beauty of profound integration of mathematics and art. With the help of fractal geometry, the design works present a kind of beauty with variation in the rules, which gives the viewer a sense of visual hierarchy and depth.

Different from the traditional color processing method, the color separation design algorithm can control the color change precisely by mathematical function, which makes the color transition more natural and harmonious. The RGB color model plays a crucial role in this process. These adjustments are usually completed through automated algorithms, thus avoiding errors and inaccuracies that may occur during manual color adjustment. This color control method based on numerical iteration not only improves the accuracy and efficiency of the design, but also gives the design work a stronger visual impact and three-dimensional sense, greatly enhancing the layering and dynamics of the design.

Under the self similarity and recursive properties of fractal geometry, design works can exhibit a unique sense of rhythm and hierarchy. The details of each fractal level are closely integrated with the overall structure, and designers use mathematical fine-tuning to precisely control each detail, so that the colors and forms of different areas echo each other, thus forming a highly logical and artistic visual effect. By continuously optimizing mathematical parameters, designers can not only achieve more precise color matching and form construction, but also harmoniously integrate every detail into the whole, enhancing the expressiveness and visual appeal of the work.

Table 1. RGB Color Modes and Numerical Range

Color component	Numeric Range	Describe
R	0-255	Red component
G	0-255	Green component
B	0-255	Blue component

The combination of fractal geometry and complex network theory provides strong mathematical support for visual design, further improving the accuracy and efficiency of graphic and color processing. Designers can create works with a sense of hierarchy, logic, and artistry through the control of algorithms, and the fine control of colors further

Table 2. RGB values (comparison table between hexadecimal and decimal)

Number	0	1	2	3	4
Decimal RGB values (R, G, B)	(0,0,0)	(0,0,252)	(36,252,36)	(0,252,252)	(252,20,20)
Hexadecimal RGB values	00 00 00	00 00 FC	24 FC 24	00 FC FC	FC 14 14

enhances the visual effect and artistic charm of the works. Therefore, the color separation design method breaks many limitations of traditional design and opens up broader possibilities for future visual creation.

By using the RGB color modes and numerical ranges listed in Table 1, designers can clearly understand the numerical ranges of each basic color in the RGB model, further guiding precise color control; Tables 2 and 3 show the correspondence between RGB values in hexadecimal and decimal, providing designers with convenient reference when adjusting colors; Tables 4 and 5 respectively list the correspondence between EGA numbers and color attributes in hexadecimal and decimal, providing necessary support for color selection and design decisions.

Table 3. RGB values (comparison table between hexadecimal and decimal continued)

Number	5	6	7	8	9
Decimal RGB values (R, G, B)	(176,0,252)	(112,72,0)	(196,196,196)	(52,52,52)	(0,0,112)
Hexadecimal RGB values	B0 00 FC	70 48 00	C4 C4 C4	34 34 34	00 00 70

The combination of fractal geometry and complex network theory provides powerful mathematical support for modern visual design, promoting the refinement of form and color processing. Fractal geometry generates complex structures with self similarity through iterative mathematical functions, which visually present rich layers and details. By combining RGB and CMYK color modes, designers can adjust the brightness, saturation, and contrast of colors under numerical control, achieving precise color matching and transitions. The additive color principle of RGB mode provides a richer color selection for digital images for screen display, while CMYK mode is widely used in the printing field to adjust color concentration according to subtractive color method. The combination of the two ensures that changes in form can be expressed in visual design to optimize the presentation of color.

In the combined application of fractal and color separation design algorithm, the designer makes the form and color echo each other in the visual space by precisely adjusting the algorithm parameters, so as to enhance the expressive force and artistic effect of the work. The designer selects the appropriate fractal formula to generate the preliminary graph, and then adjusts the gradient and layer of the color through the color separation algorithm on this basis. Through the comprehensive optimization of form and color, the design works visually present a stronger sense of three-dimensional

Table 4. EGA Number and Color Attributes in hexadecimal

EGA Number	Color property	Hexadecimal R value	Hexadecimal G value	Hexadecimal B value
EGA0	Black	00	00	00
EGA1	Blue	00	00	FC
EGA2	Green	24	FC	24
EGA3	Cyan	00	FC	FC
EGA4	Red	FC	14	14
EGA5	Magenta	B0	00	FC
EGA20	Brown	70	48	00
EGA7	White	C4	C4	C4
EGA56	Gray	34	34	34
EGA57	Lt Blue	00	00	70

Table 5. EGA number and color attribute in decimal system

EGA Number	Decimal RGB values (R, G, B)	Color property
EGA0	(0, 0, 0)	Black
EGA1	(0, 0, 252)	Blue
EGA2	(36, 252, 36)	Green
EGA3	(0, 252, 252)	Cyan
EGA4	(252, 20, 20)	Red
EGA5	(176, 0, 252)	Magenta
EGA20	(112, 72, 0)	Brown
EGA7	(196, 196, 196)	White
EGA56	(52, 52, 52)	Gray
EGA57	(0, 0, 112)	Lt Blue

and hierarchical sense. This design method based on numerical iteration improves the creation efficiency, provides designers with higher freedom, promotes the innovative development of visual design, and meets the demand for efficiency and precision in modern design.

4.2 Integration of Design Algorithms and Philosophical Aesthetics

The algorithm breaks the limitations of traditional design forms by precisely controlling every parameter in the design process, thus giving the design works a new visual experience and unique aesthetic characteristics, especially in fractal design. It gradually

evolves from simple geometric forms to complex structures through recursive iteration. In this process, the design not only shows the self-similarity that is prevalent in nature, but also creates a visual effect with rich layers and rhythm through exquisite composition. In addition, the color separation design algorithm makes the use of colors more diverse and rich through precise mathematical control of color changes, and can present a gradual process from simple to complex, from low saturation to high contrast, which makes the design works have a stronger sense of dynamics and depth, thereby enhancing the overall visual impact.

The core feature of the fractal design algorithm lies in its repeated iterative calculation process, which transforms the initial simple geometric elements into a highly complex and self-similar graphic structure. In the design, all details are continuously evolved through this iterative process, and the local and the whole always maintain internal coordination and consistency, and finally form a flowing visual rhythm. At the same time, in the color processing process of fractal design, the change of color follows similar mathematical logic. With the deepening of iteration, the visual attributes such as color contrast and saturation will gradually change, forming a gradual beauty from simple to complex and from shallow to deep, breaking the fixed pattern in traditional design methods and giving the works unique aesthetic expression. Design works can not only create stunning visual effects through fractal design algorithms, but also trigger profound thinking about the relationship between nature, order and chaos at the philosophical level. This thinking not only integrates mathematics and art more closely, but also makes the works reach a new height in terms of aesthetics. The combination of fractal geometry and complex network theory brings new aesthetic expressions to visual space design, especially in the presentation of winding beauty and broken beauty. The winding beauty comes from the self-similarity and iterative process of fractals. It simulates the tortuous forms of growth and evolution in nature to show the continuous changes of time and space in nature. These two aesthetic features enrich the expressive force of visual space through the application of fractal algorithm and bring profound philosophical significance to design. They reflect the dialectical relationship between order and disorder, continuity and fracture in nature and further expand the artistic boundary of modern design.

Ancient philosophy's understanding of nature and modern fractal geometry theory have similar internal logic. In Taoism, the viewpoint of "Tao produces one, life produces two, two produces three, and three produces all things" emphasizes the process of creation and evolution of all things in the universe, revealing the natural law from simple to complex. This idea echoes the principles of self-similarity and recursive iteration in fractal geometry, suggesting that complex structures in nature arise from the constant repetition and evolution of simple rules. "Tai Chi produces two instruments, and two instruments produce four images" in Zhouyi also emphasizes that the generation law of all things from simple to complex is consistent with the hierarchical structure and self-similarity in fractal theory.

With the development of digital technology, fractal aesthetics has been more accurately expressed in design. Designers can simulate the evolution of natural forms by computer to show the static beauty of nature, but also to show its dynamic changes. This process not only brings new creative space for visual arts, but also promotes a deeper

understanding and exploration of natural laws by humans, opening up a new perspective on the integration of art and technology.

5 Conclusion and Prospect

With the continuous advancement of computer technology, the application prospects of fractal design algorithms in the field of visual arts are becoming increasingly broad. The fractal design algorithm and color separation design algorithm proposed in this article fill the research gap in color design in current fractal art, and present unique aesthetic features through the fractal graphics generated by the algorithm. These algorithms not only provide designers with an efficient and convenient way of creation, especially suitable for designers without programming foundation, but also integrate the beauty of logic in mathematics, non Euclidean geometry, and the harmonious beauty of ancient philosophy, producing works with profound visual effects. In the future, with the further development of artificial intelligence and computer graphics, fractal design algorithms will be able to meet more complex design requirements and open up new directions for real-time interactive design and dynamic visual effects. However, how to enhance artistic expression while improving generation efficiency remains an important direction for future research. Overall, fractal design algorithms will continue to drive the development of the design field, providing more creative and personalized ways of artistic creation, and promoting the deep integration of mathematics and art.

References

1. Vaughan, J., Ostwald, M.J.: Fractal geometry in architecture. *Handbook of the Mathematics of the Arts and Sciences*. Springer International Publishing, Cham, pp 1345–1360 (2021)
2. Shi, C.: Design of fractal art design image based on one-dimensional MFDMA algorithm. In: Atiquzzaman, M., Yen, N., Zheng, X. (eds.) *Big Data Analytics for Cyber-Physical System in Smart City: BDCPS 2020*, 28–29 December 2020, Shanghai, China, pp. 247–251. Springer Singapore, Singapore (2021). https://doi.org/10.1007/978-981-33-4572-0_36
3. Yu, Z., Sohail, A., Jamil, M., et al.: Hybrid algorithm for the classification of fractal designs and images. *Fractals* **31**(10), 2340003 (2023)
4. Jahanmiri, F., Parker, D.C.: An overview of fractal geometry applied to urban planning. *Land* **11**(4), 475 (2022)
5. Lee, J.H., Ostwald, M.J.: Fractal dimension calculation and visual attention simulation: assessing the visual character of an architectural façade. *Buildings* **11**(4), 163 (2021)
6. Robles, K.E., Roberts, M., Viengkham, C., et al.: Aesthetics and psychological effects of fractal based design. *Front. Psychol.* **12**, 699962 (2021)
7. Viengkham, C., Spehar, B.: Beyond visual aesthetics: the role of fractal-scaling characteristics across the senses. *J. Perceptual Imag.* **5**, 1–14 (2022)
8. Schor, A.L., Kim, T.: A shape modulus for fractal geometry generation. *Comput. Graph. Forum* **42**(5), e14905 (2023)
9. Brielmann, A.A., Buras, N.H., Salingeros, N.A., et al.: What happens in your brain when you walk down the street? implications of architectural proportions, biophilia, and fractal geometry for urban science. *Urban Sci.* **6**(1), 3 (2022)
10. Friedenber, J., Martin, P., Uy, N., et al.: Judged beauty of fractal symmetries. *Empir. Stud. Arts* **40**(1), 100–120 (2022)



Binary Logistic Model of Smart Tourism Based on Data Information System

Danhong Chen^(✉), Lei Zhao, Yining Zhuang, Meilin Zhang, Yu Sun, and Xin Liu

School of Economics and Management, Shenyang Aerospace University, Shenyang, Liaoning, China

icatci22139@163.com

Abstract. With the rapid development of information technology, smart tourism has become a new trend in the development of tourism. Through integrated data information system, smart tourism realizes comprehensive and real-time monitoring and management of tourism activities, aiming to improve tourist experience and tourism operation efficiency. As a statistical analysis tool, binary Logistic model can effectively predict and analyze tourism-related events, which is of great significance for the construction of smart tourism system. The purpose of this study is to design a binary Logistic model of smart tourism based on data information system. Through quantitative analysis of key factors in tourism activities, tourists' behaviors and attitudes can be predicted, so as to provide decision-making support for tourism managers. This study not only helps to improve the quality and efficiency of tourism services, but also has a demonstration and promotion role for the development of smart tourism. Through the effective combination of data information system and binary Logistic model, the intelligent management and sustainable development of tourism can be realized.

Keywords: Smart Tourism · Data Information System · Binary Logistic Model · Factor Analysis · Big Data

1 Introduction

With the rapid development of the global economy and the improvement of people's living standards, tourism has become one of the important engines to promote economic growth. The growing demand for quality of travel experience is driving the travel industry to constantly seek innovation and change. In this context, smart tourism comes into being. It realizes comprehensive and real-time monitoring and management of tourism activities through integrated data information system, aiming to improve tourist experience and tourism operation efficiency. In smart tourism, data information system plays a central role. It can help tourism managers understand tourist behavior, optimize the allocation of tourism resources and improve the quality of tourism services. However, how to effectively use these data, and how to build appropriate models to predict and analyze key events in tourism activities, has become a major challenge for the current tourism industry. As a statistical analysis tool, binary Logistic model can effectively predict and

analyze tourism-related events, which is of great significance for the construction of smart tourism system. Therefore, this paper aims to design a binary Logistic model of smart tourism based on data information system to predict the behavior and attitude of tourists through quantitative analysis of key factors in tourism activities, so as to provide decision support for tourism managers.

2 Related Works

Smart tourism is a new tourism development model that uses modern information technologies, such as the Internet, big data, cloud computing and artificial intelligence, to innovate tourism services, management and operation and improve tourism experience and efficiency [1]. In the process of the development of smart tourism, not only tourism services become more intelligent and personalized, but also profound changes have taken place in the Travel mode, booking channels, information acquisition channels and consumption experience sharing of tourists, paying more attention to timeliness, convenience, individuation and customization [2]. Xu, Aristeia Kontogianni(2023)explores the role of smart travel technologies in travel planning and analyzes how travelers can use these technologies to improve travel satisfaction. The study adopts a framework of exploration and exploitation and identifies factors that promote and hinder the use of these technologie [3]. Aristeia Kontogiann iand Efthimios Alepis(2022) believes that the development of smart tourism technology not only improves the quality and efficiency of tourism services, but also provides new possibilities for the innovation and upgrading of the tourism industry, which helps to promote the sustainable and healthy development of the tourism industry [4].

Smart tourism has a profound impact on the shaping of tourism image of tourism destinations and the improvement of tourists' revisit rate, etc. Therefore, it is of great significance to explore effective means to improve tourists' satisfaction based on factor analysis and Logistic analysis. Fan Shuai (2019) conducted an empirical analysis of tourist satisfaction with factor analysis [5]. Zhang Yan (2018) adopted the binary Logistic regression method to conduct an empirical analysis on the factors affecting tourist satisfaction [6]. The above research results provide evidence for the research methods on the impact factors of smart tourism satisfaction in this paper.

In short, this paper adopts factor analysis method and Logistic regression analysis method to investigate, analyze and study the influencing factors of smart tourism tourists' satisfaction in Chengzishan from the perspective of tourists' perception evaluation, aiming to find out the basic rules of smart tourism tourists' satisfaction performance evaluation.

3 Research Methods

3.1 Database Information System

Database Information System (DBIS) is a system for organizing, storing, managing and processing large amounts of data [7]. It stores data in a structured manner in a database and provides a complete set of software tools to support data access, update, retrieval

and analysis [8]. The core group of DBIS includes the database management system sequential interface, which manages and maintains data, and the user interface, such as the query language, applications can also easily interact with the data. The main functions of DBIS include:

- Data integration: Collect and integrate data from various sources to improve the consistency and accuracy of information [9].
- Data sharing: allows multiple users or applications to access one piece of data at the same time, reducing data redundancy.
- Data security: protect data through permission control and encryption technology.
- Data analysis: Support complex data query and analysis to help enterprises make data-driven decisions [10].

3.2 Factor Analysis Method

Factor analysis is a method to reduce the total number of multidimensional variable factors on the premise of retaining the original data information as much as possible based on the correlation between variables, and strive to aggregate multidimensional variable factors into a few independent common factors and carry out weighted calculation and statistical analysis. The factor analysis model is as follows:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} \quad (1)$$

$x = Af + \varepsilon$ is the overall observation index vector.

$A = (a_{ij})$ is the factor load matrix.

F is a common factor, $F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$. The number of common factors J is called

complexity, and the smaller the value of J , the less complex the dimension of X . ε is the (other) special factor vector, which refers to other influence factors that are not included by the first J common factors.

3.3 Logistic Regression Analysis Method

Logistic regression is an important statistical method to analyze the relationship between qualitative values and the factors that contribute to the values. Since Logistic regression analysis does not require that the qualitative value and the factors leading to the value must obey the normal and the same covariance matrix, this kind of application is very wide. The binary Logistic regression model is:

$$\text{Logit}(y) = \ln \frac{p}{1-p} = a_1x_1 + a_2x_2 + \dots + a_nx_n + b \quad (2)$$

$$\frac{p}{1-p} = e^{a_1x_1+a_2x_2+\dots+a_nx_n+b} \quad (3)$$

$$p(Y = 1|x) = \frac{1}{1 + e^{-(\omega^T x + b)}} \quad (4)$$

$\frac{p}{1-p}$ is the advantage ratio, that is, the probability ratio of event occurrence to non-occurrence.

The value of $\text{Logit}(y)$ is $(-\infty, +\infty)$.

As $\text{Logit}(y)$ approaches positive infinity, the probability of $p = p(y = 1)$ gets closer.

4 Experimental Results

4.1 Selection of Experimental Indicators

At present, the evaluation index system of smart tourism is not mature enough. In order to determine the research variables, this paper draws a lot of reference from the relevant evaluation index system and constructs the evaluation index of smart tourism tourists' satisfaction (see Table 1).

Table 1. Evaluation index of smart tourism tourists' satisfaction

Primary indicator	Secondary indicator
F1 Dining Accommodation Experience	A1 dining accommodation A2 Staff service attitude A3 Dining accommodation prices A4 Overall evaluation of dishes
F2 Tourism construction and development ²	B1 sign plate B2 Public Toilet B3 Leisure facilities B4 scenic road B5 Parking Lot B6 mobile signal B7 Convenient transportation B8 route design B9 Tourism commodity development B10 Sales of tourist goods B11 Development of natural landscape
F3 Smart Travel Experience	C1 Smart travel route planning C2 tourism information intelligent real-time push C3 intelligent voice guide C4 convenient online ordering of travel products C5 Smart tourism information publicity and promotion

4.2 Experimental Results

This paper takes Chengzishan of Xifeng County, a smart tourism resource, as the research sample. Chengzishan Scenic Spot, situated in the southeast of Xifeng County, is among the first batch of cultural relics protection units and AAA scenic spots in Liaoning Province. Based on the evaluation data of tourists’ satisfaction with smart tourism in Chengzishan, 473 valid samples were collected in this paper. The T-test results of the sample data are presented in Table 2. All the T-values are negative, indicating that the actual experience and expectations of the residents at the tourist destination are consistent, and the sample data is valid.

Table 2. Sample data T test table

Variable Satisfaction	Standard deviation Satisfaction	Significance standard deviation	Average deviation	T-value	sig. (2-tailed)
A1	1.018	0.717	−0.644	−6.782	0.000
A2	0.916	0.916	−0.668	−6.115	0.000
A3	1.044	0.970	−0.712	−6.620	0.000
A4	1.097	0.793	−0.332	−3.312	0.001
B1	1.080	0.866	−0.605	−5.917	0.000
B2	0.988	0.929	−0.580	−5.766	0.000
B3	1.113	0.836	−0.454	−4.350	0.000
B4	1.143	0.832	−0.507	−4.884	0.000
B5	1.095	0.865	−0.259	−2.576	0.011
B6	1.148	0.943	−0.385	−3.610	0.000
B7	1.115	0.888	−0.283	−2.636	0.009
B8	1.148	0.860	−0.459	−4.312	0.000
B9	1.143	0.842	−0.332	−3.011	0.003
B10	1.154	0.862	−0.595	−5.617	0.000
B11	1.160	0.899	−0.337	−3.104	0.002
C1	1.176	0.879	−0.444	−4.149	0.000
C2	1.142	0.916	−0.341	−3.291	0.001
C3	1.127	0.856	−0.537	−5.024	0.000
C4	1.139	0.840	−0.546	−5.421	0.000
C5	1.181	0.834	−0.380	−3.424	0.001

5 Experimental Analysis

5.1 Model Validation of Factor Analysis

Factor analysis is a method to reduce the dimensionality of factors according to the correlation between variables, retain the original data information as much as possible, aggregate multidimensional variables into a few independent common factors, and carry out weighted calculation statistical analysis. In this paper, SPSS26.0 is used to factor the data.

1) Calculation of factor load coefficient

The factor load coefficient shows the correlation between the factor and the measured item. The measured variable passes the significance test. When the standardized load coefficient value is greater than 0.6, it indicates that the measured variable meets the requirements of factor analysis. In the results of Table 3, all the measured items reached

Table 3. Table of factor load coefficients

Factor	Variable	non-standard load factor	standardized load factor	Z	S.E	P
F1	A1	1	0.784	-	-	-
	A2	1.080	0.798	8.905	0.203	0.000***
	A3	1.703	0.747	8.738	0.195	0.000***
	A4	2.002	0.793	8.889	0.225	0.000***
F2	B1	1	0.781	-	-	-
	B2	0.901	0.785	19	0.047	
	B3	1.008	0.824	20.26	0.050	0.000***
	B4	0.926	0.750	17.939	0.052	0.000***
	B5	0.985	0.799	19.462	0.051	0.000***
	B6	0.924	0.777	18.759	0.049	0.000***
	B7	0.893	0.767	18.456	0.048	0.000***
	B8	0.933	0.758	18.185	0.051	0.000***
	B9	0.929	0.767	18.476	0.050	0.000***
	B10	1.010	0.805	19.631	0.051	0.000***
	B11	0.964	0.764	18.361	0.053	0.000***
F3	C1	1	0.785	-	-	-
	C2	0.992	0.787	19.228	0.052	0.000***
	C3	1.065	0.818	20.232	0.053	0.000***
	C4	0.977	0.792	19.369	0.050	0.000***
	C5	1.019	0.802	19.704	0.052	0.000***

the tertiary significance test and the standardized load coefficient values were greater than 0.7, meeting the requirements for further factor analysis.

2) *Model fitting evaluation*

Model fitting is to analyze the relationship between independent variables and dependent variables using approximation criteria. As shown in Table 4, the feedback results in the table all meet the judgment criteria, indicating that the model is perfect.

Table 4. Model fitting evaluation table

Common index	p	Chi-square freedom ratio	GFI	RMSEA	RMR	CFI	NFI	NNFI
Judging standard	>0.05	<3	>0.9	<0.10	<0.05	>0.9	>0.9	>0.9
Calculated value	0	1.958	0.958	0.045	0.024	0.979	0.958	0.976

3) *Analysis of covariance*

Covariance analysis is a process of interaction between covariables and factors. The closer the standard coefficient value of covariance analysis is to 1, it indicates that there is a strong correlation between factors. The data results in Table 5 show that the standard estimated coefficients of covariance of F1, F2 and F3 factors are all greater than 0.98, indicating that there is a strong correlation between the two factors, and the factor hypothesis is established, which is suitable for exploratory factor analysis.

Table 5. Factor covariance table

Factor A	Factor B	Non-standard estimation coefficient	Standard error	z	p	Standard estimation coefficient
F1	F2	0.542	0.046	11.752	0.000***	0.99
F1	F3	0.583	0.049	11.898	0.000***	0.991
F2	F3	0.595	0.051	11.736	0.000***	0.988

4) Factor weight analysis

The factor weights of F1, F2 and F3 were obtained by calculating the post-rotation variance explanation rate and cumulative variance explanation rate (see Table 6).

Table 6. Factor weight analysi

Name	Explanation rate of variance after rotation	Cumulative variance explanation rate after rotation	weight
F1	0.245	0.245	35.226%
F2	0.241	0.486	34.559%
F3	0.21	0.696	30.215%

5) Measurement of satisfaction

In this paper, the weighted average method is used to measure satisfaction, and the following formula is adopted:

$$LTU = \text{Sum}(W_i \cdot X_i), (i \in [1, 3])(i \text{ is a positive integer}) \quad (5)$$

Among them, LTU is the overall satisfaction index of tourists, W_i is the first evaluation index, and X_i is the tourists' evaluation of the i index.

Based on the weighting of F1, F2 and F3 evaluation indicators of smart tourism tourists by Chengzishan in Table 7 above, an overall satisfaction formula is obtained:

$$\text{Total LTU} = 0.3523 \cdot F1 + 0.3456 \cdot F2 + 0.3021 \cdot F3 \quad (6)$$

F1 has a weight of 35.226%, F2 34.559% and F3 30.215%. This indicates that F1 is the focus of the work to improve the overall satisfaction of tourists in the future, but other indicators of F2 and F3 factors that have a greater impact on satisfaction cannot be ignored. Therefore, this paper carries out logistic regression analysis and analyzes 20 indicators one by one.

5.2 Model Verification of Binary Logistic Regression Analysis

This paper assumes that overall satisfaction Y can be determined by a binary categorical variable. The overall experience of tourists during smart tourism is divided into two parts: the final choice of "satisfied" and "very satisfied" is recorded as 1; The final selection of "average", "unsatisfactory" and "very unsatisfactory" is marked as 0. Considering that the dependent variable of the data is a binary categorical variable, this paper adopts binary logistic regression processing and analysis. Using SPSS26.0, binary logistic regression processing information of 473 sample data was summarized as follows:

Binary logistic model regression results and analysis. The parameter regression results of the model are shown in Table 7. Among the 20 variables, 5 variables are

significant at the 10% statistical level, namely B3 leisure facilities, B4 scenic road conditions, B9 types and characteristics of tourism commodities, B11 natural landscape development and C3 intelligent audio tour, indicating that the above 5 variables are highly correlated with the overall satisfaction and need to be greatly improved.

Table 7. Parameter regression results of the model

Term	Regression coefficient	Standard error	Wald	P-value	OR value	OR value 95% confidence interval	
						<i>Upper</i>	<i>lower limits</i>
Constant	7.22	0.946	58.294	0.000***	1365.808	214.042	8715.27
A1	0.266	0.229	1.353	0.245	1.305	0.833	2.043
A2	0.263	0.211	1.554	0.212	1.301	0.86	1.968
A3	0.016	0.184	0.007	0.933	1.016	0.708	1.457
A4	0.079	0.195	0.164	0.685	1.082	0.738	1.587
B1	−0.032	0.21	0.024	0.878	0.968	0.642	1.461
B2	−0.315	0.197	2.555	0.110	0.73	0.496	1.074
B3	−0.376	0.183	4.196	0.041**	0.687	0.479	0.984
B4	−0.451	0.197	5.254	0.022**	0.637	0.433	0.937
B5	0.053	0.207	0.065	0.798	1.054	0.703	1.581
B6	−0.211	0.203	1.082	0.298	0.809	0.543	1.206
B7	−0.215	0.202	1.141	0.285	0.806	0.543	1.197
B8	−0.082	0.195	0.179	0.673	0.921	0.629	1.349
B9	−0.608	0.19	10.29	0.001***	0.544	0.375	0.789
B10	−0.003	0.19	0	0.986	0.997	0.687	1.447
B11	−0.331	0.193	2.928	0.087*	0.718	0.492	1.049
C1	−0.054	0.194	0.079	0.779	0.947	0.648	1.384
C2	0.12	0.198	0.37	0.543	1.128	0.765	1.661
C3	−0.48	0.207	5.361	0.021**	0.619	0.412	0.929
C4	0.054	0.202	0.07	0.791	1.055	0.71	1.567
C5	0.205	0.207	0.99	0.320	1.228	0.819	1.841

For the evaluation of logistic model regression results, the closer the AUC value is to 1, the better the classification effect is. As can be seen in Table, the five data of logistic model regression result evaluation are all close to 1, which indicates that the model regression result is reasonable and the conclusion is valid.

6 Conclusion

Through the above factor analysis and Logistic regression analysis, this paper believes that the factors affecting the satisfaction of Chengzishan smart tourism tourists include catering and accommodation experience (F1), leisure and recreation facilities (B3), scenic road (B4), tourism commodity development (B9), natural landscape development (B11) and intelligent audio tour (C3). Therefore, combining with the weak links of Chengzishan smart tourism learned from further interview and investigation, it is suggested to take effective measures to improve them. For example, improve the accommodation experience, promote food and beverage, build recreation facilities, open up transportation and tourism, create scenic IP, develop tourism products, and strengthen intelligent voice guidance.

Acknowledgment. This paper is supported by the research project of Economic and Social Development of Liaoning Province in 2025 (2025lslybkt-088).

References

1. Kontogianni, A., Alepis, E., Virvou, M., Patsakis, C. Smart Tourism-The Impact of Artificial Intelligence and Blockchain. *Intelligent Systems Reference Library* 249, pp. 1–178. Springer 2024 (2023)
2. Kontogianni, A., Alepis, E.: Tourism embraces blockchain towards the smart tourism era. *Intell. Decis. Technol.* **17**(3), 811–838 (2023)
3. Kontogianni, A.: User Crowdsourced and Crowdsensed Data and Artificial Intelligence Enhanced Mobile apps for Smart Tourism and Smart Cities. University of Piraeus, Greece (2023)
4. Kontogianni, A., Alepis, E.: Social Network Data Enabling Smart Tourism. *IISA* 2023 1–6 (2022)
5. Shuai, F.: Research on tourist satisfaction in rural tourism destinations based on factor analysis and IPA analysis. *J. Kaifeng Educ.* **39**(06), 292–293 (2019)
6. Zhang, Y., Liu, J.: Research on tourist satisfaction of natural scenery scenic spots in Beijing based on binary Logistic regression. *J. Arid Land Resour. Environ.* **32**(11), 202–208 (2018)
7. Mesgari, M., Mohajeri, K., Azad, B.: Affordances and information systems research: taking stock and moving forward. *ACM SIGMIS Database* **54**(2), 29–52 (2023)
8. Raatikainen, P., Pekkola, S., Mäkelä, M.: Narrativization in information systems development. *J. Database Manag.* **35**(1), 1–30 (2023)
9. Ning, Y.: Research on Computer Information Hotel Front-End Storage Database Integrated Management System. *CIPAE*, pp. 70–73 (2023)
10. Ru, X. Development of employee information management system based on B/S mode and SQL server database. In: *CIPAE*, pp. 245–249 (2023)

Author Index

A

An, Dongyang 508
An, Zhenpeng 265

B

Bai, Wanrong 508

C

Chang, Xiaopeng 43
Chen, Danhong 596
Chen, Haokun 508
Chen, Huimin 585
Chen, Junpeng 325
Chen, Junru 256
Chen, Lifei 346
Chen, Manjiang 162
Chen, Miao 220
Chen, Siyu 43
Chen, Wei 412
Chen, Xi 370
Chen, Xiaoning 314
Chen, Xinkai 105
Chen, Yongshu 56
Chu, Jingyi 294
Cui, Yanchao 174
Cui, Zhaoxia 184

D

Diao, Zhiqiang 508
Dong, Peng 540
Du, Junliang 325

F

Fan, Yijiao 464
Fan, Yufeng 265
Fang, Xuming 151

G

Gao, Yan 208
Gui, Tengyue 425

Guo, Jiarui 56

Guo, Mengyan 239

H

He, Guijiao 33
He, Xin 81
Hong, Haisheng 56
Huang, Kuiwen 530
Huang, Tianyou 551

J

Jia, Hanjie 162
Jiang, Dongming 496
Jiang, Hua 162

K

Ke, Hongyu 81

L

Li, Aimin 92
Li, Cuiping 184
Li, Haosheng 412
Li, Ming 151
Li, Qiang 105
Li, Shuai 496
Li, Xiangyang 346
Li, Yan 444
Li, Yongna 184
Li, Zhenting 585
Lin, Qin 56
Liu, Caihua 346
Liu, Jian 256
Liu, Shuyan 248
Liu, Xin 596
Lv, Wanting 412

M

Ma, Wei 105
Ma, Yixuan 412

Q

Qi, Wenyue 476

R

Ren, Qingqing 412

S

Shang, Dachao 401

Shao, Xinru 174

Shen, Xiaoying 174

Shi, Bingjiao 357

Shi, Chunhe 174

Shi, Ximei 304

Song, Junyan 1

Su, Feng 346

Sun, Nuan 174

Sun, Yu 596

Sun, Zheng 56

T

Tan, Qing 453

Tang, Huajun 530

Tang, Xiaoning 127

Tang, Yi 81

W

Wang, Dong 346

Wang, Erlan 573

Wang, Husong 425

Wang, Jieru 485

Wang, Junliang 346

Wang, Lei 13

Wang, Liying 140

Wang, Qi 519

Wang, Xiaoyi 325

Wang, Xiwen 24

Wang, Ya 239

Wang, Yaran 174

Wang, Yimin 357

Wang, Yiqun 508

Wang, Yishan 275

Wang, Yuxin 561

Wei, Jianmei 115

Wei, Wenting 70

X

Xi, Xiuliang 115

Xie, Mucun 401

Xu, Hanyue 283

Xu, Haobin 425

Xu, Meng 140

Xu, Ning 81

Xu, Shu 573

Xu, Shuting 435

Xu, Weimin 425

Xu, Zixiang 275

Y

Yang, Chenming 92

Yang, Jinzhu 380

Yang, Shuo 81

Yang, Xinlei 151

Yao, Lei 229

Yao, Li 334

Yao, Yuan 105

Yu, Bo 43

Yu, Haowen 530

Yuan, Taiping 530

Z

Zeng, Shuyan 401

Zhai, Shangyu 508

Zhang, Bangcheng 43

Zhang, Fan 239

Zhang, Haiming 573

Zhang, Jie 70

Zhang, Jiong 140

Zhang, Lei 265

Zhang, Ling 194

Zhang, Meilin 596

Zhang, Moxin 357

Zhang, Qingwang 412

Zhang, Xiyu 43

Zhang, Yinsong 208

Zhang, Zeyuan 496

Zhao, Lei 596

Zhao, Zhenghui 220

Zhao, Ziyang 325

Zheng, Yang 325

Zhou, Junlin 585

Zhou, Xiaokai 275

Zhu, Jie 530

Zhu, Longjie 151

Zhu, Pengwei 391

Zhu, Zhaoyu 81

Zhu, Zhifang 56

Zhuang, Yining 596