

Srikanta Patnaik
Madjid Tavana
Vipul Jain *Editors*

New Paradigms in Big Data Technology and Business Analytics

Proceedings of the International
Conference on Big Data Technology and
Business Analytics

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).

Srikanta Patnaik · Madjid Tavana · Vipul Jain
Editors

New Paradigms in Big Data Technology and Business Analytics

Proceedings of the International Conference on
Big Data Technology and Business Analytics

Editors

Srikanta Patnaik
Interscience Institute of Management
and Technology
Bhubaneswar, Odisha, India

Madjid Tavana
La Salle University
Philadelphia, PA, USA

Vipul Jain
Victoria University of Wellington
Wellington, New Zealand

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-96652-1

ISBN 978-3-031-96653-8 (eBook)

<https://doi.org/10.1007/978-3-031-96653-8>

© 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

Big Data Technology and Business Analytics are transforming the way organizations operate in the digital era. By collecting, processing, and analyzing vast volumes of structured and unstructured data, businesses can uncover valuable insights that drive strategic decisions. Big data platforms such as Hadoop, Spark, and cloud-based solutions enable efficient data storage and real-time processing. These technologies support predictive analytics, allowing companies to forecast market trends, customer behavior, and operational risks. Business analytics tools leverage these datasets to create dashboards, reports, and visualizations that aid in understanding complex scenarios. As a result, companies can optimize resource allocation, improve customer satisfaction, and enhance overall performance. Advanced analytics techniques like machine learning and natural language processing further expand the potential of big data in identifying hidden patterns and anomalies. Industries such as finance, healthcare, retail, and manufacturing are increasingly adopting these technologies to stay competitive. However, successful implementation requires robust data governance, skilled professionals, and a clear strategy aligned with business goals. Ultimately, Big Data Technology and Business Analytics empower organizations to become more agile, informed, and innovation-driven.

The primary aim of the International Conference on Big Data Technology & Business Analytics (BDTBA-2024) is to bring together a global community of scholars, researchers, industry experts, and practitioners to discuss and share advancements in big data technologies and business analytics. The conference serves as a vibrant platform for exchanging innovative ideas, presenting cutting-edge research, and exploring real-world applications of data-driven solutions. In today's rapidly evolving digital landscape, the integration of big data and analytics has become essential for informed decision-making, operational efficiency, and competitive advantage. BDTBA-2024 is committed to fostering interdisciplinary dialogue, encouraging collaborative research, and promoting the development of novel tools and frameworks that address complex business challenges. The conference provides an opportunity to examine the impact of artificial intelligence, machine learning, cloud computing, and other emerging technologies on data analysis and business strategies. Participants engage in meaningful discussions that bridge the gap between academic research and industry practices. The event also highlights the importance of ethical data use, data governance, and security in the era of massive information flow. By hosting keynote speeches, technical sessions, and panel discussions, BDTBA-2024 ensures a comprehensive learning experience for all attendees. It aims to inspire innovation and support the next generation of data professionals and researchers. With great enthusiasm, we present the conference proceedings, showcasing the rich diversity of thought and knowledge shared at BDTBA-2024.

Big Data Technology and Business Analytics hold a futuristic vision of transforming the way we understand, interpret, and act upon information in every sector of society. As data becomes the new driving force of innovation, these technologies enable smarter decision-making, personalized services, and predictive insights that can revolutionize

industries such as healthcare, education, finance, and public governance. BDTBA-2024 embraces this vision by creating a collaborative platform where experts can exchange groundbreaking research and practical solutions. The conference encourages the development of responsible, scalable, and efficient data practices that address real-world societal challenges. Through its diverse sessions and knowledge-sharing initiatives, BDTBA-2024 aims to provide inclusive growth, technological empowerment, and data literacy. Ultimately, the conference contributes to building a data-driven society that is more informed, resilient, and sustainable.

Srikanta Patnaik

Editorial

Big Data Technology and Business Analytics is to utilize vast and complex datasets to extract meaningful insights that support strategic decision-making and innovation. By leveraging advanced tools and techniques, such as machine learning, data mining, and real-time analytics, organizations can gain a competitive edge in a rapidly evolving digital landscape. This field aims to improve efficiency, forecast trends, and optimize performance across various domains, including healthcare, finance, manufacturing, and education. The futuristic approach of big data and analytics envisions intelligent systems that can autonomously analyze data, adapt to new information, and support proactive decision-making. As data continues to grow exponentially, the focus shifts toward scalable architectures, ethical data use, and enhanced data governance. Together, these advancements pave the way for a smarter, more connected, and data-driven future.

The International Conference on Big Data Technology & Business Analytics (BDTBA-2024), hosted by Interscience Institute of Management and Technology, Bhubaneswar, India, during November 22nd–23rd, 2024. This conference provided an exceptional platform for researchers, researchers, academics, practitioners, and industry experts to submit their original research papers, case studies, and practical experiences for presentation at the conference. BDTBA-2024 aims to provide a premier platform for knowledge exchange, collaboration, and innovation in the fields of Big Data Technology and Business Analytics.

Big Data Technology and Business Analytics are profoundly reshaping today's world by enabling faster, smarter, and more data-driven decisions across industries. From improving healthcare outcomes to optimizing business operations, their influence is evident in every aspect of modern life. These technologies also lay a strong foundation for the next generation, equipping them with tools and insights to innovate, solve complex problems, and build a more efficient and informed society. BDTBA-2024 witnessed a focused exploration of these themes, with the organizing committee receiving 128 submissions; after rigorous review and editing, 51 papers were accepted for presentation and publication, of which 40 papers were officially registered and are included in this proceeding. These contributions are categorized into five major sections: (i) Big Data Infrastructure and Technologies, (ii) Data Mining and Machine Learning, (iii) Big Data Applications and Case Studies, (iv) Business Intelligence and Decision Support, and (v) Data Governance and Ethics.

Big Data Infrastructure and Technologies

The first track of BDTBA-2024, *Big Data Infrastructure and Technologies* consist of five articles, focuses on the foundational systems and frameworks that support the effective handling and processing of large-scale data. This track highlights the importance of robust, scalable, and secure infrastructure to enable real-time analytics and informed

decision-making across sectors. The article *“Overview of Research on Compliance Testing of Data Standards in the Transportation Sector”* addresses the critical need for standardized data protocols to ensure interoperability and efficiency in transportation systems. *“Design and Implementation of a Supply Chain Traceability System Based on Blockchain Technology”* explores how blockchain enhances transparency and security in supply chain management. Another significant contribution, *“A Data Synchronization Desensitization System Based on Format-Preserving Encryption,”* presents a novel approach to safeguarding sensitive data while maintaining operational accuracy. *“Research on the Development Strategy of Smart Tourism Driven by Big Data Platform”* showcases how big data infrastructure can revolutionize tourism experiences through personalized services and intelligent planning. The paper *“Construction of Digital Twins Model for Logistics Finance”* demonstrates the integration of digital twin technology with big data to optimize logistics and financial processes. Collectively, these articles reflect the evolving landscape of big data infrastructure and its critical role in driving digital transformation across industries.

Data Mining and Machine Learning

The second track of BDTBA-2024, *Data Mining and Machine Learning*, centers on the extraction of meaningful patterns from vast datasets and the development of intelligent models to drive predictive and prescriptive analytics. This track contains ten selective articles which emphasizes how machine learning techniques are revolutionizing decision-making across industries by enabling systems to learn from data and improve performance over time. The article *“The Study of Financial Risk Analysis and Prediction for Enterprises Based on Big Data Mining”* illustrates the powerful application of data mining in assessing potential financial threats and developing strategies to mitigate them. *“Mean-CVaR Portfolio Model Based on Graph Neural Network”* explores an innovative approach in financial analytics by combining neural networks with risk-sensitive portfolio optimization. The paper *“Correlation Analysis of Corporate Culture and Employee Behavior: A Predictive Model Based on Deep Learning”* showcases the role of deep learning in uncovering behavioral patterns that influence organizational outcomes. Meanwhile, *“Big Data Mining Techniques and Their Optimisation in Financial Data Analysis”* focuses on the enhancement of data mining algorithms for better accuracy and efficiency in financial systems. A key contribution to operational forecasting is *“Research on the Application of LSTM-Prophet Model in Cargo Volume Prediction of E-Commerce Logistics Sorting Center,”* which combines advanced time-series models to enhance logistics planning. Similarly, *“Deep Learning in Multi-scale Analysis of Financial Time Series”* applies multi-resolution analysis to uncover complex patterns in financial datasets. In the domain of cybersecurity, the article *“An Improved RFS-XGBoost Based Model for Network Intrusion Detection System”* demonstrates a machine learning-based solution to detect and prevent network threats in real time. The paper *“Strategic Decision-Making and Risk Control in Financial Modeling Based on LSTM Artificial Intelligence Methods”* highlights the power of long short-term memory models in dynamic and uncertain decision environments. Health-focused innovation is reflected in *“Study on Machine Learning Algorithm to Optimize Functional Movement*

Intervention Strategy for the Elderly,” showcasing how intelligent systems can support personalized healthcare and rehabilitation. Finally, “*Advanced Forecasting of Financial Data with Neural Networks*” presents cutting-edge approaches to improving the accuracy of financial predictions through artificial neural networks. Together, these ten articles exemplify the breadth and depth of research in data mining and machine learning, spanning diverse fields such as finance, logistics, healthcare, and cybersecurity. Each study leverages the capabilities of modern algorithms to solve real-world problems with greater precision and efficiency. The track highlights how data-driven methodologies are shifting traditional paradigms toward smarter, automated systems. As data grows exponentially, the significance of mining it effectively and learning from it becomes crucial for competitive advantage. Machine learning continues to evolve, with deep learning, graph networks, and hybrid models pushing boundaries. These advancements foster innovation, reduce risks, and enable accurate forecasting. Moreover, the ethical implications of machine learning are increasingly being addressed through robust models and responsible data handling practices. The papers included in this track serve as key exemplars of both academic rigor and practical relevance. BDTBA-2024 showcases how the synergy between data mining and machine learning is shaping a smarter and more connected future.

Big Data Applications and Case Studies

The third track of BDTBA-2024, Big Data Applications and Case Studies, highlights how advanced data analytics are being practically implemented across a range of industries and social sectors. This section showcases 14 insightful articles that demonstrate the versatility and real-world value of big data. In the field of education and communication, the study on utilizing big data analysis to improve English cross-cultural communication ability presents innovative ways to enhance intercultural interactions. In transportation and infrastructure, the traffic congestion prediction model for smart freeways illustrates how big data can drive smart city solutions. Financial risk is another key theme, with works such as the application of big data analysis technology in financial risk identification and comparative analysis of FinTech risks in China and the USA offering valuable insights for banking and finance sectors. Commerce and industry also benefit from data-driven approaches, as reflected in the articles on e-commerce platform innovation, cross-border supply chain development, and product purchasing process optimization. The impact of digital transformation on regions like Guangdong’s manufacturing sector highlights the connection between technology and sustainability. Articles such as the AI-based forecasting model for international trade commodity demand and prediction of agricultural consumer preferences bring attention to how data enhances market intelligence. The track also includes critical discussions on modern enterprise economic management and educational strategy optimization through information technology. Collectively, these case studies underscore the tangible benefits of big data applications in improving decision-making, enhancing efficiency, and fostering innovation across diverse sectors. This track exemplifies how data is no longer confined to theoretical modeling but is actively transforming real-world practices and strategies.

Business Intelligence and Decision Support

The fourth track of BDTBA-2024, *Business Intelligence and Decision Support*, delves into advanced methods for enhancing strategic business decisions through data-driven insights. This track features five thought-provoking articles that explore the intersection of analytics, finance, and management. The study on *precision marketing for special agricultural product e-commerce* demonstrates how user profiling and key technologies can revolutionize targeted business strategies. An *intelligent financial risk identification and early warning model* powered by Big Data Platforms (BDPs) showcases proactive measures for risk mitigation. The track also includes a compelling analysis of *earnings management amid digital transformation*, using statistical tools like Stata to evaluate corporate performance. Further, insights on *engineering project decision support systems* and *empirical evaluations of diversification strategies* round out the track, offering practical frameworks for informed and intelligent decision-making in modern enterprises.

Data Governance and Ethics

The concluding track of BDTBA-2024, *Data Governance and Ethics*, presents a forward-looking examination of how digital innovations intersect with ethical considerations and organizational integrity. This segment features six compelling articles that highlight the broader societal and managerial implications of technology-driven transformation. One study explores how *enterprise digital transformation* can drive energy efficiency, aligning innovation with sustainability. Another paper delves into the *infiltration of digital finance* and its psychological impact on consumption behavior, raising important ethical and socio-economic questions. The role of *information technology in the informatization of management accounting* is examined to reveal shifts in transparency and accountability in the digital age. A thoughtful exploration of *human resource management systems rooted in integrity culture* emphasizes the importance of values in digital system design. The track also includes research on *virtual reality's impact on corporate culture experience*, showcasing immersive technology's potential in shaping ethical organizational environments. Lastly, a *computer-aided study on dual-employee family work engagement* offers insights into how digital tools can inform inclusive workplace strategies grounded in ethical resource use and human-centered design.

Big Data Technology and Business Analytics continue to evolve as transformative forces across industries, revolutionizing how information is collected, interpreted, and utilized. Despite their immense potential, challenges such as data privacy concerns, integration complexities, and the need for real-time processing remain at the forefront. Addressing these issues requires robust infrastructure, ethical frameworks, and innovative methodologies. The futuristic scope of this field includes enhanced automation, predictive intelligence, and AI-driven decision systems that will redefine operational excellence. As organizations increasingly adopt data-centric approaches, the demand for skilled professionals and interdisciplinary research will surge. The *International Conference on Big Data Technology & Business Analytics (BDTBA-2024)* serves as a vibrant platform to address these evolving demands. It aims to bring together researchers,

practitioners, and thought leaders to exchange cutting-edge ideas and present impactful case studies. The objective is to foster collaboration, inspire innovation, and chart a sustainable and intelligent future powered by big data. With a global outlook and a multidisciplinary focus, the conference emphasizes both theoretical advancements and real-world applications. As we conclude, BDTBA-2024 stands as a testimony to the growing significance of data as the backbone of future decision-making and societal development.

Acknowledgement

We would like to express our sincere appreciation and gratitude to all those who have contributed to the success of the International Conference on Big Data Technology & Business Analytics (BDTBA-2024).

First of all, we would like to express our heartfelt thanks to all the contributors of *BDTBA-2024*, who have generously shared their research findings and insights through the papers compiled in this volume. Their active involvement, thought-provoking discussions, and valuable contributions have made the conference a rich and intellectually stimulating experience. The papers presented in this conference explore critical themes such as Big Data Infrastructure and Technologies, Data Mining and Machine Learning, Big Data Applications and Case Studies, Business Intelligence and Decision Support and Data Governance and Ethics. The conference emphasizes the importance of utilizing data analytics and technological innovation to address global challenges, support sustainable development, and empower the next generation of digital transformation. Through these academic engagements, *BDTBA-2024* has set a strong foundation for continued research and collaboration in the realm of big data and business analytics. The ideas, findings, and synergies generated during this event are sure to leave a lasting impact, propelling forward the vision of a data-informed, intelligent society.

Our deepest appreciation goes to the Program and Technical Committees for their diligent work in reviewing and selecting high-quality papers. Their scholarly expertise and unwavering commitment to academic rigor ensured a program of exceptional depth and relevance. We also extend our sincere gratitude to the *Organizing Committee* for their dedication, meticulous planning, and seamless coordination, which contributed immensely to the successful execution of *BDTBA-2024*.

We are especially thankful to the *Springer Series Editors of Lecture Notes in Networks and Systems*-**Janusz Kacprzyk** for his continued support in publishing the conference proceedings “*New Paradigms in Big Data Technology and Business Analytics*.” in the esteemed book series. Our special thanks also go to **Thomas Ditzinger**, *Editorial Director at Springer*, for his guidance and cooperation throughout the publication process.

We are sincerely grateful to all individuals and organizations who played a role in making *BDTBA-2024* a successful and memorable event. Your support and commitment have been instrumental in realizing the conference’s vision. Thank you all for your valuable contributions toward shaping the future of Big Data Technology and Business Analytics.

Srikanta Patnaik
Madjid Tavana
Vipul Jain

BDTBA-2024 Conference Committee

General Chair

Madjid Tavana

Business Systems and Analytics Department La
Salle University Philadelphia, Pennsylvania,
USA

General Co-chair

Florin Popențiu Vlădicescu

University “Politehnica” in Bucharest, Romanian

Program Chair

Srikanta Patnaik

IIMT, Bhubaneswar, Odisha, India

Program Co-chair

E. Ertugrul Karsak

Galatasaray University, Turkey

Publication Chair

Vipul Jain

Victoria University of Wellington, New Zealand

Publication Co-chairs

Yusuf Joseph Ugras

La Salle University Philadelphia, USA

Organizing Chairs

Ajit Kumar Nayak

Department of Computer Science and Information
Technology Siksha 'O' Anusandhan Deemed
to be University, Bhubaneswar

Organizing Co-chair

Wu Chun Ho Jack

The Hang Seng University of Hong Kong, China

Conference Convener

Mrs. Soma Mitra

IIMT, Bhubaneswar, Odisha, India

Technical Program Committee

Indranil Bose

IIM Ahmedabad, India

U Dinesh Kumar

IIM, Bangalore, India

Sobhan Sarkar

IIM, Ranchi, India

Praveen Ranjan Srivastava

IIM, Rohtak, India

Bodhibrata Nag

IIM, Calcutta, India

Mahamaya Mohanty

IIM, Rohtak, India

Josyula Srinivas

IIM, Vishakapatnam, India

Sumeet Gupta

IIM, Raipur, India

Rahul Kumar

IIM, Sambalpur, India

Pragati Priyadarshinee

IIM, Bodhgaya, India

Suresh Kumar Jakhar

IIM, Lucknow, India

Sanjeet Singh

IIM, Lucknow, India

S. Venkataramanaiah

IIM, Lucknow, India

Bharat Bhasker

IIM, Lucknow, India

Omkarprasad S Vaidya

IIM, Lucknow, India

M. P. Sebastian

IIM, Kozhikode, India

Vandana, IIM

Sirmaur, India

Vinu C T, IIM

Tiruchirapally, India

Gajanand M. S.

IIM, Tiruchirapally, India

Bunil Kumar Balabantaray

NIT, Meghalaya, India

P. S. Apirajitha

Anna University, Chennai, India

S. Kanimozhi

VIT University, Chennai, India

Saswati Mukherjee

Anna University, Chennai, India

N. B. Muthu Selvan	Sri Sivasubramaniya Madar College of Engineering, Chennai, India
R. Seyezhai	Sri Sivasubramaniya Madar College of Engineering, Chennai, India
L. Sai Ramesh	Anna University, Chennai, India
T. T Mirmalinee	SSN College of Engineering, Chennai, India
Chitta Ranjan Mallick	PMEC, Berhampur, Odisha, India
Jyoti Ranjan Nayak	ICFAI, Hyderabad, India
Ramamani Tripathy	Lovely Professional University
Rudra Kalyan Nayak	VIT, Bhopal, India
Dulu Patnaik, GEC	Kalahandi, Odisha, India
Jude Hemanta	karunya University, India
Mitali Madhusmita Nayak	Siksha 'O' Anusandhan University, India
Brij Mohan Gupta	IIT, Delhi, India
Rabinarayan Satpathy	Sri Sri University, India
Vipul Jain	Victoria University of Wellington, New Zealand
Suchana Mishra	Dayananda Sagar College of Engineering, Bengaluru, India
Bijayananda Patnaik	NIT, Raipur, India
Trilok Saine, DEAL	DRDO, Dehradun, India
Madhumita Panda	GMU, Sambalpur, Odisha, India
Chandra Sekhar Panda	Sambalpur University, Odisha, India
Rajat Kumar Panigrahi	ITI, Berhampur, Odisha, India.

International Advisory Committee

Sabyasachi Patnaik	F.M. University, Balasore, Odisha, India
A. Ridha Mahjoub	University Paris Dauphine, France
Ali Emrouznejad	Aston University, UK
Briance Mascarenhas	Rutgers University-Camden, USA
Farookh Khadeer Hussain	University of Technology Sydney, Australia
Francisco Javier Santos Arteaga	Free University of Bolzano, Italy
Gholam R. Amin	University of New Brunswick, Canada
Gregory Ulferts	University of Detroit Mercy, USA
Gül Temur	Bahcesehir University, Turkey
Hasan Bal	Gazi University, Turkey
Hür Bersam Bolat	Istanbul Technical University, Turkey
Ibrahim H. Osman	American University of Beirut, Lebanon
Imed Kacem	Université de Lorraine, Nancy, France
Jana Hančlova	VŠB-Technical University of Ostrava, Czech Republic

Joshua Ignatius	University of Warwick, United Kingdom
Josef Jablonský	University of Economics Prague, Czech Republic
José Luis Verdegay	University of Granada, Spain
Long Pham	University of Louisiana, USA
Luisa Tibiletti	University of Torino, Italy
Mariana Rodrigues de Almeida	Universidade Federal do Rio Grande do Norte, Brazil
Matteo Rocca	Università degli Studi dell'Insubria, Italy
Mehtap Dursun	Galatasaray University, Turkey
Michael Patriksson	Chalmers University of Technology, Sweden
Mohsen Attaran	California State University-Bakersfield, USA
Rosa Lombardi	Sapienza University of Rome, Italy
Sahand Daneshvar	Eastern Mediterranean University, Turkey
Simone Domenico Scagnelli	Edith Cowan University, Australia
Sungmook Lim	Dongguk University, Korea
Vincent Charles	University of Buckingham, UK

Contents

Big Data Infrastructure and Technologies

Overview of Research on Compliance Testing of Data Standards
in the Transportation Sector 3
Jing Deng

Design and Implementation of a Supply Chain Traceability System Based
on Blockchain Technology 13
Lushi Liuai, Yiling Wu, and Zhaoyun Han

A Data Synchronization Desensitization System Based
on Format-Preserving Encryption 23
Jingao Chen, Xiaoxu Hu, Peng Zhang, Yu Guo, and Yunjun Ma

Research on the Development Strategy of Smart Tourism Driven by Big
Data Platform 33
Yuehua Bai and Yu Yang

Construction of Digital Twins Model for Logistics Finance 44
Guangren Chen

Data Mining and Machine Learning

The Study of Financial Risk Analysis and Prediction for Enterprises Based
on Big Data Mining 57
Min Jiang

Mean-CVaR Portfolio Model Based on Graph Neural Network 67
Guogang Zheng

Correlation Analysis of Corporate Culture and Employee Behavior:
A Predictive Model Based on Deep Learning 80
Meichao Song and Li Gao

Big Data Mining Techniques and Their Optimisation in Financial Data
Analysis 91
Qingle Chen

Research on the Application of LSTM-Prophet Model in Cargo Volume Prediction of E-Commerce Logistics Sorting Center	103
<i>Yufei Huang, Zhijie Huang, Haoyun Yang, and Meishan Chen</i>	
Deep Learning in Multi-scale Analysis of Financial Time Series	114
<i>Nannan Sun</i>	
An Improved RFS-XGBoost Based Model for Network Intrusion Detection System	123
<i>Weiyi Liu</i>	
Strategic Decision-Making and Risk Control in Financial Modeling Based on LSTM Artificial Intelligence Methods	133
<i>Zhuqi Wang, Zhuopei Cheng, and Qinghe Zhang</i>	
Study on Machine Learning Algorithm to Optimize Functional Movement Intervention Strategy for the Elderly	143
<i>Huijun Wang</i>	
Advanced Forecasting of Financial Data with Neural Networks	156
<i>Yuanpeng Song and Diansheng Yang</i>	
Big Data Applications and Case Studies	
A Study on Utilizing Big Data Analysis to Improve English Cross-Cultural Communication Ability	171
<i>Rongrong Ye</i>	
Application of Big Data Analysis Technology in Financial Risk Identification	181
<i>Zhifeng Qu</i>	
Traffic Congestion Prediction Model for Smart Freeways Based on Big Data Analysis	193
<i>Zhengxuan Wen</i>	
Development of an AI-Based International Trade Commodity Demand Forecasting Model	207
<i>Wenle Qin, Yi Ling Wu, and Nan Chen</i>	
Research on Optimization Method of Product Purchasing Process Based on ASME Analysis——Take QH Company as an Example	217
<i>Chunjie Li and Jinfeng Wen</i>	

Digital Transformation and Sustainable Development of Guangdong's Manufacturing Sector	232
<i>Guiling Fan</i>	
Research on the Application of Information Technology in Modern Enterprise Economic Management	244
<i>Fanling Kong and Qiwen Zhang</i>	
The Impact of Fintech Development on Commercial Banks' Credit Risk: The Case of SPD Bank	256
<i>Xu Ruoyu, Wang Heng, and Xu Yanni</i>	
Exploring the Avenues for Optimizing University Education Management Through the Empowerment of Information Technology	265
<i>Shuqing Wang</i>	
Research on the Development and Innovation Strategies of E-commerce Platforms Driven by Big Data Technology	277
<i>Ruodan Luo, Ke Chen, and Chao Wang</i>	
Research on Prediction Model of Agricultural Product Consumer Preferences Based on Big Data Analysis	288
<i>Ziyu Zhang</i>	
Research on Supply Chain Development Strategy of Cross-Border E-commerce Enterprises Empowered by Digital Technology	298
<i>Lihua Yang</i>	
Comparative Analysis of FinTech Risks in China and the U.S.: An Application of the GARCH-VaR Model	307
<i>Ming Zhao and Meiling Du</i>	
A Study of the Time-Varying Shock Effects of Economic Policy Uncertainty on the Prices of Small Agricultural Commodities——Garlic and Ginger as an Example	318
<i>Jingjing Chen and Guixian Tian</i>	
Business Intelligence and Decision Support	
Research on Key Technologies and Innovative Strategies for Precision Marketing of Special Agricultural Product E-commerce Based on User Profile	331
<i>Zhenyan Hu and Qiongjie Zhou</i>	

Research on Intelligent Financial Risk Identification and Early Warning Model Driven by BDP	343
<i>Xiaodi Fan and Ge Guo</i>	
Analyzing the Impact of Digital Transformation on Earnings Management Based on Stata Software: Taking H Company as an Example	353
<i>Zhijia Zhang and Jiaqi Ma</i>	
Decision Support System for Engineering Project Management Based on Data Mining	363
<i>Dayong Zhang and Rongwen Chen</i>	
Performance Analysis of Diversification Strategies in Internet Service Enterprises: An Empirical Study	374
<i>Li Zhou</i>	
Data Governance and Ethics	
Research on the Impact and Innovation Strategies of Enterprise Digital Transformation on Energy Utilization Efficiency	385
<i>Juanjuan Tian and Ruiqin Tang</i>	
Research Progress on the Infiltration of Digital Finance and Relative Deprivation of Consumption	397
<i>Kenjie Chen, Shaohui Zhang, and Chuyuan Jiang</i>	
Analysis of the Impact of Information Technology on the Informatization of Management Accounting in the Digital Economy Era	406
<i>Fanling Kong and Shuyan Meng</i>	
Research on the Information System Construction of Enterprise Human Resource Management Based on Integrity Culture	415
<i>Weiling Hou and Lili Xie</i>	
Innovative Application of Virtual Reality Technology in Corporate Culture Experience	425
<i>Jing Wang and Ning Yang</i>	
Analysis on the Influencing Factors of Dual-Employee Family Work Engagement Based on Resource Conservation Theory: A Computer Aided Study	434
<i>Xiaorui Shi and Yiwen Chen</i>	
Author Index	445

About the Editors

Srikanta Patnaik is the Director of IIMT, Bhubaneswar. He has received his Ph. D. (Engineering) on Computational Intelligence from Jadavpur University, India, in 1999 and supervised 34 Ph. D. theses and more than 60 M. Tech theses in the area of Machine Intelligence, Soft Computing Applications and Re-Engineering. He has published more than 100 research papers in international journals and conference proceedings. He is author of 2 textbooks and edited 82 books and few invited book chapters, published by leading international publisher like Springer-Verlag, Kluwer Academic, etc. He is the Editors-in-Chief of International Journal of Information and Communication Technology and International Journal of Computational Vision and Robotics published from Inderscience Publishing House, England and also Editors-in-chief of Book Series on “Modeling and Optimization in Science and Technology” published from Springer, Germany.

Madjid Tavana is a Professor and Distinguished Chair of Business Analytics at La Salle University, where he serves as Chairman of the Business Systems and Analytics Department. He also holds an Honorary Professorship in Business Information Systems at the University of Paderborn in Germany. He is Distinguished Research Fellow at the Kennedy Space Center, the Johnson Space Center, the Naval Research Laboratory at Stennis Space Center, and the Air Force Research Laboratory. He was recently honored with the prestigious Space Act Award by NASA. He holds an MBA, PMIS, and PhD in Management Information Systems and received his Post-Doctoral Diploma in Strategic Information Systems from the Wharton School at the University of Pennsylvania. He has published 22 books and over 300 research papers in international scholarly academic journals. Tavana is the Editor-in-Chief of Decision Analytics Journal, Healthcare Analytics, Space Mission Planning and Operations, International Journal of Applied Decision Sciences, International Journal of Management and Decision Making, International Journal of Communication Networks and Distributed Systems, and International Journal of Knowledge Engineering and Data Mining. He is also an editor of Annals of Operations Research, Expert Systems with Applications, Computers and Industrial Engineering, Intelligent Systems with Applications, and Journal of Innovation and Knowledge.

Vipul Jain is a Senior Lecturer in Operations and Supply Chain Management at Victoria Business School, Victoria University of Wellington, New Zealand. He has also worked as a French Government researcher for the French National Institute for Research in Computer Science and Control at Nancy, France, and was also involved in the European project I* PROMS (Innovative Production Machines and Systems) from UK. He has more than 85 archival publications to his credit in high impact factor journals, as well as conference papers and books chapters, is the Editor-in-Chief of International Journal of Intelligent Enterprise, and an Editorial Board member for seven international journals. He is ranked 7th in India in a list of top 20 Indian leading academicians publishing

in the area of Logistics and Supply Chain management in an exhaustive study entitled “Analysis of the Logistics research in India- white paper (2012), pp.1-11” conducted by researchers from Institute of Transport Logistics, Dortmund University, Institute for Supply Chain and Network Management, Technische Universität Darmstadt, and University of Münster, Germany. He is involved as a scientific and advisory committee member for various international conferences. He was the Program Coordinator for the Masters in Industrial Engineering Program at IIT Delhi, and has been actively involved in conducting Management Development Programs/Training Programs for Senior Managers and Executives of Public and Private Organizations in India on various issues and challenges encountered in operations, logistics, and supply chain management.

Big Data Infrastructure and Technologies



Overview of Research on Compliance Testing of Data Standards in the Transportation Sector

Jing Deng^(✉)

Resource Center China Academy of Transportation Sciences, Beijing 100029, China
jingdeng1994@163.com

Abstract. Informatization, digitalization and intelligence are key characteristics of high-quality development in the transportation industry, and data elements are the new driving forces for the development of transport. Data standardization is a fundamental task in the development of digital transportation, and it is crucial for ensuring that information systems achieve information sharing, interconnection and business collaboration. The compliance testing of data standards promotes data standardization and quality management, which is of great significance for promoting data fusion applications and business system collaboration and linkage. This paper reviews the current state of compliance testing both domestically and internationally and summarizes the key aspects of data standard compliance testing, including the subjects, content, processes, standards and platform tools. It analyzes the current state of compliance testing research in the transportation industry and offers development suggestions to support the optimization and upgrading of data standard compliance testing work in the transportation industry information systems.

Keywords: Data Elements · Standard Compliance · Transportation · Compliance Testing · Testing Platform Tools

1 Introduction

Data is a key production factor in the digital economy era, and an important engine for improving the efficiency of digital development, expanding the functions of integrated development, and increasing the momentum of business innovation in transportation. With the rapid development and integration of new-generation information technology, the transportation industry, using data as a key element, is accelerating the construction of key informatization projects and demonstration pilot projects. The development levels of transportation informatization, digitalization, and intelligence continue to improve. Data resources have become the new driving force for transportation development, information systems have penetrated all business areas of the transportation industry, and the industry has entered a new stage of comprehensive networking, business collaboration, and intelligent application.

Despite significant advances in informatization and digitalization within the sector, issues such as weak-data-foundations and poor-application-synergy still exist [1]. In the

important strategic opportunity period of innovation-led and data-enabled transformation and upgrading of traditional industries, there is a serious contradiction between abundant data and lack of information in the transportation industry [2]. Currently, the information construction of the national transportation is in a critical period of transformation from fragmentation to integration and from isolation to open sharing. The construction of a strong transportation nation places higher demands on the development of intelligent transportation systems [3]. Adhering to the “Single Platform” Perspective and enhancing the integration and sharing, comprehensive development, and intelligent application of data resources are critical measures to improve the digital level of the industry, promote high-quality development of comprehensive transportation and accelerate the construction of a strong transportation nation [4].

Perfecting the data quality assurance and open sharing system and building a standardized high-quality transportation data resource “sharing pool”, can effectively improve the supply level of transportation data elements, and effectively solve the problems such as severe fragmentation of information systems, difficulties in interconnectivity and business collaboration. Data standardization is a fundamental task in information construction, serving as a technical precursor for data sharing and system collaboration. Under the planning and guidance of the Transportation Industry Informatization Standard System, the transportation industry has introduced over a hundred informatization data standards in recent years. However, the standard specification system for the whole life cycle of data collection, integration, storage, processing and transaction has not yet been formed, the actual implementation effect of the standard is different, and the implementation effect of the standard lacks follow-up supervision means, and the data quality and standardization level still need further improved.

Standard compliance testing is the use of testing means to verify whether the information system conforms to the standards and the degree of compliance and supervise the implementation and application of standards in information projects. Data standard compliance testing is based on testing technologies to assess the compliance of data items and data sets within information systems to standards, effectively ensuring data quality and interoperability, supporting the implementation and monitoring of standards, and promoting the interconnection and resource sharing of information systems. In December 2023, the National Data Administration along with seventeen other departments jointly issued the Three-Year Action Plan (2024–2026) for ‘Data Elements X’, which in the aspect of Data Elements X Transportation, proposes to advance the construction and reuse of high-quality data sets, and promote the fusion application of multi-source data. The release of the data multiplier effect will drive the synergistic optimization of data elements, enhance reuse efficiency, and foster innovative integration, thereby promoting the development of new productive forces while also imposing higher demands on data quality and standardization [5]. Strengthening data standard compliance testing, standardizing quality management throughout the entire lifecycle of data, is of great significance for fully leveraging the value of data, promoting data fusion applications, and business transformation and upgrading.

2 Data Standard Compliance Testing

Data-class standards regulate the identification, content, classification, and formatting of data during the processes of data collection, processing, and transmission, ensuring data standardization and quality, and supporting data exchange sharing and interconnectivity of information systems. Standard compliance testing began earlier internationally, with foreign standardization organizations having developed over nine hundred standards and technical specifications related to compliance testing. Research institutions have also introduced various testing tools, including platforms for SQL compliance testing.

Domestically, there is less research related to standard compliance testing, but as informatization progresses, various domestic industry sectors have begun to emphasize and advance related work on industry standard compliance testing. Currently, data standard compliance testing in China has seen explorations in research and application in sectors such as auditing, education, healthcare, transportation, media publishing, and international trade. For instance, in the education sector, the E21 Education Management Informatization Standard Intelligent Testing Analysis System primarily tests database field definitions and codes, while in the auditing sector, the National Standard Compliance Testing System for Accounting Software Data Interfaces can implement data control for software interconnectivity [6, 7].

2.1 Testing Subjects and Content

Data standard compliance testing should be conducted at the design, implementation, and acceptance stages of information systems. The main subjects of testing include the data dictionary, the data table structure, and the data content, with specific test items being data type, data format, data unit, data range and other attributes, as well as consistency and integrity.

(1) Feasibility Study and Design Phase

The feasibility study report for an information construction project should clearly propose the standards the system must follow; preliminary design documents should specify the data to be exchanged and shared by the system, organized by data table or data items; during the detailed design process, the data dictionary should be compiled in accordance with standard specifications. At this stage, the data dictionary document should undergo standard compliance testing.

(2) Implementation Phase

During the project construction implementation phase, compliance testing should be performed on the system's data dictionary and data table structure and the database should be modified and perfected based on the test report.

(3) Acceptance Phase

After the system has been tested, run, and accumulated a certain amount of data, compliance testing should be conducted on the data dictionary, data table structure, and data content. The test report should be submitted at the time of acceptance.

2.2 Testing Process

Data standard compliance testing can be divided into the following three main stages, as shown in Table 1:

(1) Test Preparation

Conduct standard parsing, clarify specific requirements for coding objects and code tables, and classify the code [8]. Simultaneously perform data preprocessing to standardize the data dictionary, database table structure, and data item attributes of the system under test. In addition, it is necessary to build a comprehensive and clear data standard compliance testing standard library based on the system situation and testing requirements.

(2) Testing Execution

Conduct compliance testing according to the testing plan, using intelligent algorithms to associate data with standards, testing attributes such as type, format, unit, and value range of the data to be inspected, as well as the consistency and integrity of the data table. The test results are judged as conformance, weak conformance and non-conformance and form a test record.

(3) Testing Results Analysis

Establish evaluation rules, analyze the test results comprehensively, calculate the overall compliance rate of the system, evaluate the standard compliance of the system being tested and generate a test report [9].

Table 1. Data standard compliance testing process tables.

Stage	Sub Stage	Description
Test Preparation	Standard Analysis	Identify all coding objects and code tables, and classify the codes
	Data Preprocessing	Normalize the data dictionary, data table structure, and attributes of data items of the system under test
	Standard Library Construction	Build a comprehensive and clearly tiered standard library
Testing Execution	Intelligent association	Based on intelligent algorithms to associate data and standards
	Data attribute testing	Testing the type, format, unit, value range, and other attributes of the data to be inspected
	Data table testing	Testing the consistency, integrity, and other contents of the data table
	Result determination	Determine the test results as compliant, weakly compliant, or non-compliant, and form a test record

(continued)

Table 1. (continued)

Stage	Sub Stage	Description
Testing Results Analysis	Evaluation rules	establish evaluation rules and comprehensively analyze data testing results
	Compliance rate calculation	Calculate the overall compliance rate of the system under test
	Evaluation and Reporting	Evaluate the compliance of the data standards of the system under inspection and generate a test report

2.3 Testing Standards

To enhance the reusability and consistency of standard compliance testing technologies, standardization organizations both domestically and internationally actively develop compliance testing specifications, standardizing the methods, tools, and evaluations involved.

Testing the Compliance of an Equipment with Its Reliability, Maintainability and Availability Specifications (CEPT T/N 45-01 E-1988) is one of the earliest internationally referenced standards mentioning compliance testing. Since the 21st century, the establishment and application of compliance testing standards have noticeably accelerated, with over a hundred compliance testing standards released internationally, covering sectors such as agriculture, construction, energy, and finance [10].

The development of compliance testing standards in China has been relatively slow, with more than half of the current compliance testing standards being based on international standards. However, certain independent standards have been released in fields such as information technology, though the application of compliance testing methods is not comprehensive enough, and the quality of testing standards still needs to be enhanced. Currently, China has developed over seventy compliance testing standards, with rapid progress in the military sector's compliance testing. In the area of data standard compliance testing, in 2021, the Requirements of standard conformance testing for information classification coding and metadata (GB/T 41139-2021) was issued, which specifies the testing methods, content, basis, results determination, and documentation requirements for data standard compliance testing.

2.4 Testing Platforms and Tools

Supporting the entire standard compliance testing process, testing platforms and tools provide functions such as data import/export, automatic analysis and processing, and report generation, effectively enhancing the automation level of testing, and improving the efficiency and quality of tests. In order to adapt to the continuous improvement, updating, and development of information technology standards in relevant countries

and industries, and to meet the requirements of high-quality and efficient standard compliance testing, the data standard compliance testing system meets the characteristics of scalability, accuracy, standardization, and high performance in overall design.

According to testing objectives and processes, testing platforms generally include modules such as data processing, compliance detection, standard library management, and evaluation and analysis. The functional modules are schematically represented as shown in the diagram below Fig. 1.

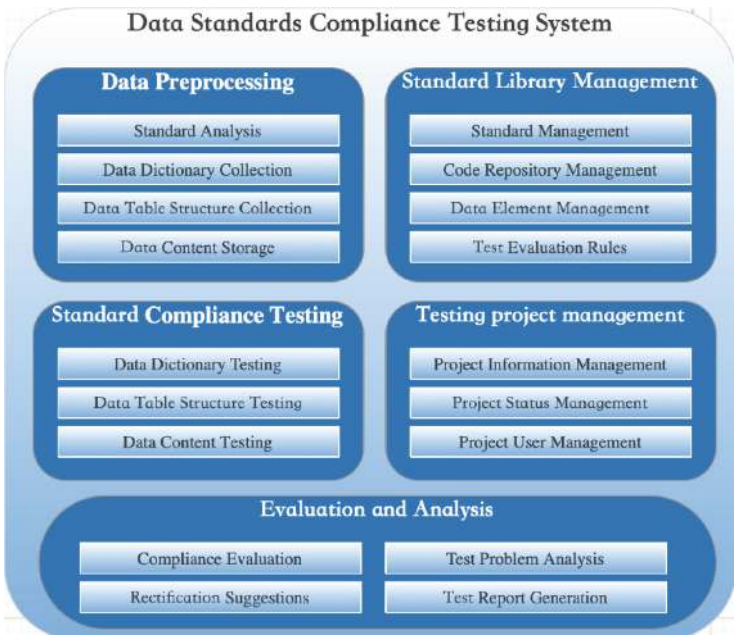


Fig. 1. Functional Structure Diagram of Testing Platform Tools

Data Preprocessing Module: Preprocesses standards and the data of systems to be tested, mainly including standard analysis, collection of user data dictionaries, collection of database structures, and tokenization of data content for storage.

Compliance Testing Module: Conducts standard compliance testing, mainly involving the testing of data dictionary documents, data table structures, and data content.

Standard Library Management Module: Establishes testing and evaluation rules, and manages the information of the standard library, primarily including standard management, code repository management, and data element management.

Testing Project Management Module: Manages the information of testing projects, including project information, project status, testing processes, and associated users, through information management systems.

Evaluation and Analysis Module: Evaluates the compliance of the system under test based on evaluation rules, analyzes the content that failed to meet the scores, provides rectification suggestions, and automatically generates test reports.

The workflow of the data standard compliance testing system mainly consists of two parts: basic data management, testing, and result analysis. When conducting testing work, first create a new project and project users for the system to be tested. Follow the process of selecting criteria, launching, uploading inventory files, matching criteria, compliance testing, and auditing to perform standard compliance testing in sequence, achieving comprehensive evaluation of system data standard compliance and exporting test results. The testing process of the data standard compliance testing system is shown in Fig. 2.

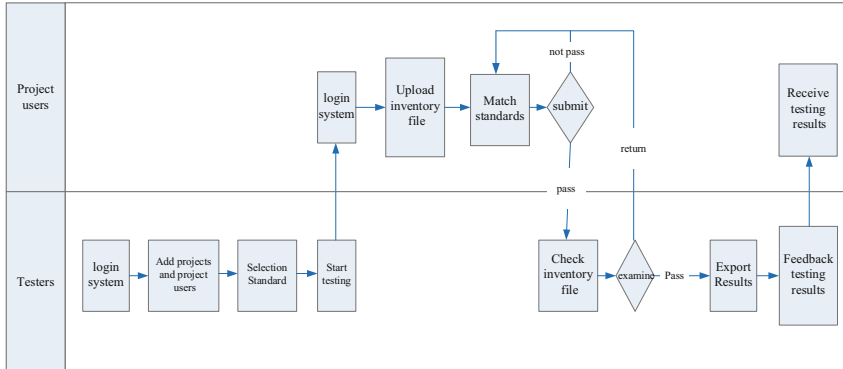


Fig. 2. Test flowchart of data standard compliance testing system

3 Current Status of Data Standard Compliance Testing in the Transportation Sector

Currently, the transportation industry has published over one hundred information resource standards, forming a relatively comprehensive data standard system. However, the development of data standards employs a top-down management approach, which has shown deficiencies in overall planning, drafting, and management [11].

Since 2011, the transportation industry has initiated research projects on foundational transportation information standards, compliance testing key technologies and specifications, leading to the creation of data element hierarchical models and a Standards Compliance Testing System for transportation information data. Based on this research, the industry standard Transportation Information System-Data Dictionary Specification(JT/T 1020–2016) was developed, which specifies the content, domain meanings, and filling requirements for the data dictionary of transportation information systems; also, the Shaanxi Provincial Standard Transportation Data Compliance Test Specification (DB 61/T 1479-2021) was formed, defining terms, general provisions, testing subjects and content, and testing and judgment criteria for transportation data compliance testing. An industry standard Transportation Information System – Data Compliance Test (JT/T 1489–2024) has been established, which specifies the overall requirements for compliance testing between transportation information system data and standards, testing methods and results determination, as well as subsequent testing requirements.

Furthermore, in 2013, the Ministry of Transportation's Department of Science and Technology issued the Technical Regulations for Testing Compliance of Traffic Information Data Standards (Draft for Comment), but it was not formally promoted or applied subsequently. In 2017, the National Transportation and Logistics Public Information Platform Management Center established a standard compliance testing system, conducting compliance testing on the connected platforms, application systems, and software. In 2019, the China Academy of Transportation Science conducted research on data standard compliance testing for transportation industry information systems, building a standard library for standard compliance testing and rules for data definition parsing in multi-type database systems, according to the testing process of four stages: data preprocessing, data element matching, compliance testing, and compliance evaluation, designing and implementing a Standard Compliance Testing System for the transportation industry, conducted a comprehensive evaluation of data standard compliance, and based on the comprehensive evaluation score, formed four levels of judgment results A-B-C-D, enhancing the standardization and effectiveness of compliance testing.

Research on data standard compliance testing in the transportation industry's information systems is still in its initial stages; testing methods and software tools require further study, and there is a lack of unified standards for testing, lacking professional theoretical and technical support in practical testing operations, and thus unable to effectively meet the needs for data standard compliance testing before the completion and acceptance of transportation informatization systems.

4 Development Recommendations for Data Standard Compliance Testing in the Transportation Sector

Standard Compliance Testing is one of the crucial means to ensure the standardization of data resources during information system construction, laying the foundation for interconnectivity and data exchange within information systems. Currently, the level of data standardization in transportation information systems is not high, and there are various degrees of issues with data standard compliance at different stages of the system lifecycle. To further improve the data quality of industry information systems, and to support system interconnectivity and business collaboration, the following areas should be further optimized in the field of transportation data standard compliance testing.

(1) Accelerate the Development and Compilation of Data Standard Compliance Testing Standards.

To ensure the practical application of data standards, strengthen top-level design, and promptly establish methods for compliance testing, as well as national and industry standards for standard database construction.

(2) Release Convenient Data Standard Retrieval and Compliance Testing Tools

Construct a standard database and provide public services for standard information retrieval based on informatization platforms, to expand the awareness and coverage of standards. Develop convenient automated tools for Standard Compliance Testing to facilitate self-testing of data standard compliance by project development and construction entities, supporting the effective advancement of third-party Standard Compliance Testing.

- (3) **Strengthen Data Standard Compliance Review During the Design Phase**
Enhance data standard compliance review during the system design phase, implement early intervention testing principles, and produce compliance testing reports for databases and data dictionaries, which are essential documents for system construction and acceptance.
- (4) **Enhance the Management of Institutions Conducting Standard Compliance Testing for Information Systems.**
Standardize the entry requirements for testing institutions and manage them through unified registration. Regularly train and assess the personnel of testing institutions to improve the level of testing services.

5 Conclusion and Outlook

Conducting data standard compliance testing on information systems can effectively enhance the enforcement of data standards, improve the quality and standardization of system data, and is of great significance in supporting the standardized construction of information systems, open data sharing applications, and business collaboration. Standard Compliance Testing is crucial for improving the supply level of data elements, strengthening data governance, and enhancing integration applications. Currently, domestic research on data standard compliance is relatively sparse, and the work on data standard compliance testing in the transportation sector urgently needs optimization. This paper re-views the current state of data standard compliance research both domestically and internationally, summarizes the subjects, processes, and platform tools of data standard compliance testing, analyzes the current state of standard compliance testing research in the transportation industry, and proposes development recommendations.

Acknowledgment. This work has been supported by Basal Research Fund of Central Public Research Institute of China (Grant No. 20190208 and No. 20240203), the Construction Project of China Knowledge Centre for Engineering Sciences and Technology (CKCEST-2022-002001).

References

1. Geng, J.L.: Information technology construction of transportation under the perspective of 'Internet+'. *Transport Bus. China* **22**, 64–66 (2023)
2. Yin, R.H., Yao, Z.F.: Research on data quality assessment framework for multidisciplinary standards. *Standard Sci.* **01**, 92–95 (2020)
3. Xue, H.P., Yang, H.J.: Analysis of China's transportation informatization construction under the background of big data. *FORLAND* **01**, 189–191 (2024)
4. Wu, Z.H., Wu, X.B., et al.: Prospect of development trend of smart transportation under the background of building china into a country with strong transportation network. *Transport Res.* **5**(04), 26–36 (2019)
5. Wang, J.L.: Research on value and industry development of IOV data elements. *Inform. Commun. Technol. Policy* **50**(03), 35–38 (2024)
6. Mao, X., Xian, K.T., et al.: Survey on standards compliance testing on data context. *Standard Sci.* (07), 39–42 (2016)

7. Chen, Y., Song, W.B.: Research on evaluation index system for interchange standards compliance test. *Ship Electron. Eng.* **40**(11), 145–147+161 (2020)
8. Zhu, H., Gao, A., et al.: Research on conformance testing of data standard. *Standard Sci.* (07), 53–57(2019)
9. Zheng, P.Q.: Evaluation index system of compliance test for information business collaboration standards. *J. Fuzhou Univ. (Nat. Sci. Edition)* **49**(06), 747–752 (2021)
10. Xu, W.X., Zuo, X., et al.: Comparison and countermeasure analysis of standard compliance testing at home and abroad. *Modern Bus. Trade Industry* **25**(01), 155–157 (2013)
11. Wang, P.C., Ma, J.C., et al.: Research on the construction of basic data meta standards for transportation information. *Transport. Technol. Manag.* (13), 9–11 (2023)



Design and Implementation of a Supply Chain Traceability System Based on Blockchain Technology

Lushi Liu¹, Yiling Wu², and Zhaoyun Han³✉

¹ Guangdong University of Science and Technology, Dongguan 523083, Guangdong, China

² Hong Kong Shue Yan University, Hong Kong 999077, China

³ Guangzhou Medical University, Guangzhou 511495, Guangdong, China
13922295089@163.com

Abstract. With the increasing demand from consumers for product safety and quality, there is an urgent need for enterprises to establish a transparent and trustworthy product traceability system. In response to this need, this paper designs and implements a supply chain traceability solution based on blockchain technology. The system adopts a semi-decentralized network architecture, combining the advantages of consortium chains and public chains to achieve efficiency and openness. Through optimizations such as blockchain sharding and state trees, the system possesses high throughput and large storage capacity. The smart contract module ensures the immutable certification of critical data, while IPFS is utilized for storing large files. Additionally, the system integrates with internal enterprise systems such as ERP to enable data sharing and business coordination. Following comprehensive functional testing, performance testing, and security assessments, the traceability system meets expected metrics and can operate stably in production environments, effectively addressing pain points related to product trust and supply chain transparency.

Keywords: Blockchain · Product Traceability · Product Safety · Quality Regulation

1 Introduction

In today's consumer environment, product quality and safety are increasingly emphasized. Counterfeit products not only seriously infringe upon consumer rights but also have a significant negative impact on corporate brand image. Meanwhile, as the supply chain extends, the circulation path of products from raw material procurement to final sales becomes increasingly complex, making it more challenging to trace product origins and production processes. Therefore, establishing a secure, transparent, and reliable product traceability system has become an urgent priority. Blockchain, as an innovative distributed ledger technology, inherently possesses characteristics such as decentralization, immutability, and traceability, providing a new technological path for addressing traceability pain points. This paper introduces a novel traceability system design based on

blockchain, explaining how to leverage blockchain advantages to construct a transparent supply chain data platform, achieve trustworthy traceability throughout the product life-cycle, and share specific practices for system implementation, testing, and optimization, providing valuable insights for enterprise digital transformation.

2 System Requirements Analysis

2.1 Business Requirements Analysis

As consumers increasingly prioritize food safety and product quality, enterprises face significant challenges in enhancing supply chain transparency and ensuring product traceability. This is crucial for establishing a strong brand image and earning consumer trust. A survey of 2,000 global enterprises revealed that 67% had experienced supply chain disruptions, with an average financial loss exceeding \$1.8 million per incident. These disruptions often stem from issues such as counterfeit products and raw material quality problems, leading to substantial financial and reputational damage [1]. Thus, constructing a secure and efficient traceability system has become an urgent priority. By implementing robust traceability mechanisms, companies can monitor and verify the entire supply chain process, from raw materials to finished products, ensuring authenticity and quality. This not only mitigates risks but also enhances consumer confidence and loyalty, ultimately contributing to a more resilient and trustworthy brand. Table 1 illustrates the key components of an effective traceability system and their respective benefits.

Table 1. Cost of Supply Chain Interruptions in Different Industries (Million US Dollars)

Industry	Average Cost (Million US Dollars)
Pharmaceutical	19.8
Automotive	1.16
Consumer Goods	8.58
Electronics	8.73

2.2 Functional Requirements Analysis

Blockchain-based supply chain traceability systems must cater to the core requirements of various stakeholders. In the food industry, for instance, suppliers need to record information about agricultural product cultivation; processors must document data regarding ingredient intake and processing stages; logistics providers require real-time tracking of the transportation process; distributors and retailers need to verify the compliance and origin of products; and consumers must be able to access comprehensive product information throughout the entire lifecycle. Therefore, the system should provide reliable and traceable functionalities for every stage of the product lifecycle, ensuring

efficient multi-party queries and trust verification [2]. Additionally, for cross-border supply chain businesses, the system must include compliance auditing features that support multi-language interfaces and adherence to international regulations. This ensures that the traceability system not only enhances transparency and accountability within domestic supply chains but also addresses the complexities of global trade, fostering greater trust and compliance across different regulatory environments.

2.3 Non-functional Requirements Analysis

Blockchain-based supply chain systems encounter challenges such as high concurrency and large-scale data management, necessitating high-performance read/write capabilities. For instance, a food company selling 10 million units annually, with each product's traceability data being 1KB, would generate 1TB of data on the blockchain per year. The system must optimize throughput and storage capacity using technologies like state trees and blockchain sharding. Additionally, high availability is crucial to handle node failures, ensuring continuous system operation. Scalability is essential to accommodate business growth, while data reliability and privacy are paramount to prevent theft and tampering. The system must also support the modularization of functional components and automated operation and maintenance to reduce operating costs for enterprises. This includes ensuring that the system can easily adapt to changing business requirements, integrate new features, and maintain efficient performance, ultimately contributing to a more resilient and cost-effective supply chain management solution [3].

3 System Design

3.1 Blockchain Network Design

Considering the geographical distribution of supply chain participants, we deploy a public chain architecture for the blockchain network. Public chain networks, characterized by decentralization and openness, are well-suited for cross-organizational supply chain environments. To enhance the network topology, we combined the benefits of consortium chains and public chains to create a semi-decentralized network model. In this model, core nodes are operated by consortium members responsible for consensus and block generation, while edge nodes can be freely joined by all participants for data querying and writing. This hybrid approach leverages the efficiency of consortium chains and the openness of public chains [4].

To further improve system throughput, we implement blockchain sharding technology, dividing the network into multiple shards, each independently completing the consensus process. The shard coefficient is set to 16, theoretically increasing performance by 16 times. Each shard employs the Practical Byzantine Fault Tolerance (PBFT) consensus algorithm to ensure consistency while enhancing block generation speed. Figure 1 illustrates the logical architecture of the network, demonstrating how this design balances efficiency, security, and scalability to meet the demands of a global supply chain.

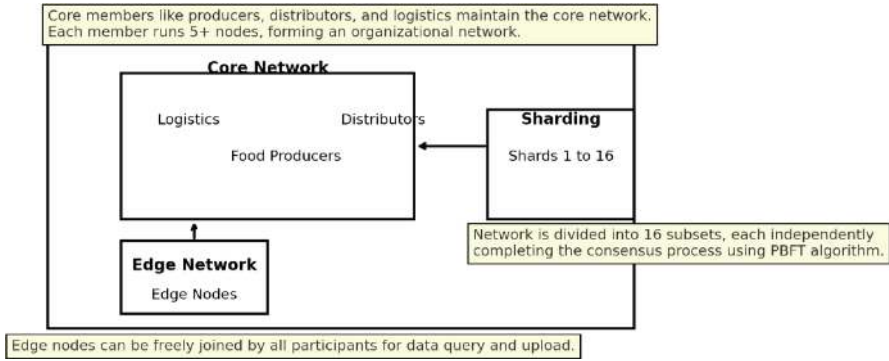


Fig. 1. Blockchain Network Topology Architecture

3.2 Smart Contract Design

Smart contracts are the core components of the blockchain-based supply chain system, responsible for recording and validating traceability data. These contracts are written in Solidity and run on the Ethereum Virtual Machine (EVM). The system comprises several key contracts, each serving a specific function:

- 1) **AssetsTracker Contract:** This contract is used to register static data of products or raw materials. It includes information such as the product name, category, initial storage location, and other pertinent details. This contract ensures that all foundational data about the assets is accurately recorded and immutable once stored on the blockchain.
- 2) **EventLogger Contract:** The EventLogger contract is designed to record key node data throughout the product's lifecycle. This includes time stamps, locations, and operator information for various activities such as processing, packaging, and both inbound and outbound logistics. By capturing these critical events, the contract provides a comprehensive and traceable history of the product's journey through the supply chain [5].
- 3) **OwnershipTransfer Contract:** This contract manages the logic for transferring product ownership. It ensures that ownership changes are conducted in compliance with established protocols, maintaining the integrity and security of the handover process. This is crucial for tracking the product as it moves from one entity to another within the supply chain.
- 4) **Access Control Contracts:** These contracts handle permission management, ensuring that only authorized users can access or modify specific data. They are essential for maintaining the security and confidentiality of the traceability information.
- 5) **Audit Contracts:** Audit contracts are used for compliance auditing. They verify that all transactions and data records adhere to the relevant regulations and standards, providing an additional layer of trust and accountability.

By implementing these smart contracts, the system ensures a reliable, transparent, and secure method for tracking products throughout the supply chain. This design not only enhances traceability and accountability but also supports efficient multi-party

queries and trust verification, ultimately contributing to improved consumer confidence and operational efficiency.

3.3 Database Design

The inherent immutability of blockchain data makes it a core storage medium for traceability data. However, to alleviate the storage pressure on the blockchain, we use IPFS (InterPlanetary File System) as auxiliary storage for multimedia data related to raw materials and products. Key node data, such as timestamps and transaction details, are stored directly on the blockchain, ensuring their security and immutability. Meanwhile, multimedia data, such as images and videos, are stored off-chain in IPFS, with only their IPFS hash fingerprints recorded on the blockchain for reference. Additionally, relevant procurement, inventory, and shipment data are maintained within enterprise internal systems, such as ERP and CRM systems. To fully leverage this heterogeneous data, we have established a centralized database that integrates with the blockchain system. This centralized database allows for the aggregation and analysis of data from various sources, providing a comprehensive view of the supply chain. By using this hybrid storage approach, we ensure the efficient handling of large-scale data while maintaining the security and integrity of critical traceability information. This design allows for the seamless integration of blockchain and traditional database systems, optimizing storage and ensuring quick access to essential data. Table 2 illustrates the distribution of data between these storage solutions, highlighting the balance between on-chain and off-chain data management:

Table 2. Traceability Data Distribution

Data Type	Storage Location
Core Traceability Data	Blockchain
Large File Data	IPFS
Enterprise Internal Data	Centralized Database

By adopting a hybrid storage architecture, the system can balance the immutable advantage of blockchain data while addressing the shortcomings of limited storage space and query efficiency on the chain [6]. It integrates seamlessly with existing enterprise systems, effectively supporting supply chain traceability applications.

4 System Implementation and Testing

4.1 System Functionality Implementation

The core functionality of the system includes data uploading, querying, and verification for traceability purposes. We adopted a B/S architecture, with the frontend developed using the React framework and the backend built on Node.js to provide RESTful API

services. Taking agricultural product traceability as an example, the data uploading process is as follows: Suppliers fill in basic information about agricultural products such as variety, planting time, location, and pesticide usage through the frontend interface. Upon submission, it triggers the registration function of the backend AssetsTracker contract, which writes this information into the blockchain as the initial traceability data of the product. During the subsequent circulation process, each participating party needs to record relevant operation data at each key node by calling methods of the EventLogger contract [7]. For example, distributors need to record receiving time, location, and operator information upon receipt, while processors need to record production dates, process parameters, etc., during processing. These key node information are immutably recorded on the blockchain, ensuring data integrity and trustworthiness. Finally, consumers can scan the QR code on the product packaging using a mobile device to view the complete information from planting to circulation, thereby identifying the product source. Table 3 illustrates the complete traceability key node data for a certain agricultural product:

Table 3. Key Node Data for Agricultural Product Traceability

Node	Time	Location	Operator	Other Information
Planting	2023/3/15	Nanjing, Jiangsu	Farmer Zhang San	Variety: Selenium-enriched rice; Use of organic fertilizers
Harvest	2023/7/10	Nanjing, Jiangsu	Farmer Zhang San	Total yield: 5000 kg
Processing	2023/8/1	Nanjing Rice Company, Jiangsu	Worker Li Si	Process: Water-cooled polishing; Packaging specification: 5 kg/bag
Logistics	2023/8/5	En route to Shanghai	Logistics Company A	Transport: Truck; Shipment volume: 20 tons
	2023/8/8	Arrival in Shanghai	Logistics Company A	
Sales	2023/8/10	Huaqian Supermarket, Shanghai	Clerk Wang Wu	Incoming quantity: 5000 bags

4.2 System Deployment and Integration

To ensure high availability and scalability of the system, we adopted a container cloud deployment model. The frontend, backend services, and blockchain nodes of the system are all packaged into Docker container images and deployed on Alibaba Cloud’s ECS container instance cluster, managed and orchestrated using Kubernetes [8]. This cloud-native microservices architecture provides the system with excellent elasticity

and scalability, allowing resources to be scaled up or down according to demand. We utilized the Hyperledger Fabric framework for the underlying blockchain network. The Fabric network is maintained by multiple organizations, including food producers, distributors, and logistics providers. Each organization runs 3–5 nodes internally to form an organizational network, and all organizational networks collectively constitute the consortium chain network. The network adopts RBFT as the consensus mechanism to ensure that transactions are not manipulated by nodes. To achieve integration with enterprise internal systems, we developed a data synchronization service that regularly retrieves relevant business data from ERP, WMS, OMS, and other systems, such as purchase orders, supplier information, inbound and outbound records, shipment tracking, etc., and synchronizes them to a MySQL database for query and usage by upper-layer applications. Figure 2 shows the growth trend of integrated data over the past year.

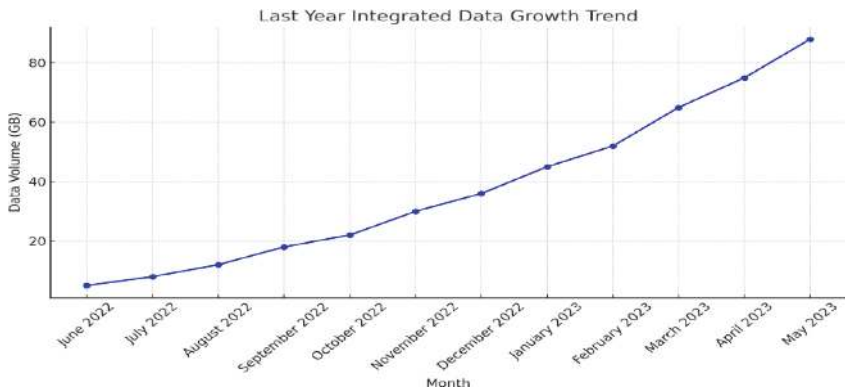


Fig. 2. Trend of Integrated Data Growth in the Past Year

With the seamless integration solution described above, the blockchain traceability system not only possesses the advantages of transparency, decentralization but also seamlessly integrates with the existing IT architecture of enterprises, maximizing data resource sharing and business coordination.

4.3 System Testing

During the system development process, we embraced the DevOps philosophy, incorporating continuous integration and continuous delivery practices to achieve rapid iteration and high-quality delivery. For functional testing, we developed numerous behavior-driven automated test cases covering core functions such as data uploading, querying, compliance auditing, and more. These test cases ensure comprehensive coverage of the system's capabilities and functionalities. We executed over 120,000 automated tests, achieving a test coverage rate of 98%. This extensive testing process ensures the correctness and integrity of the system's functionality, identifying and resolving potential issues early in the development cycle. The high test coverage also provides confidence in the system's reliability and robustness. Figure 3 illustrates the statistical distribution

of test cases, highlighting the thoroughness of our testing efforts and the areas of focus to ensure a well-functioning and secure traceability system. This rigorous testing regime is crucial for maintaining the high standards expected in a blockchain-based supply chain traceability solution [9].

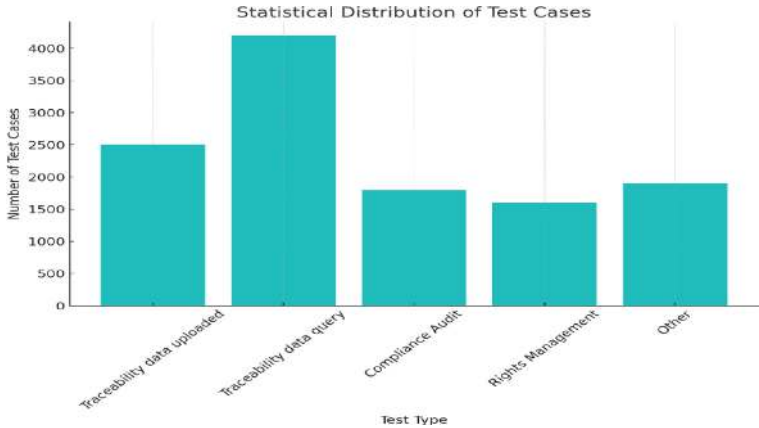


Fig. 3. Statistical Distribution of Test Cases

During the performance testing phase, we utilized tools such as Apache JMeter to simulate high-concurrency scenarios and conducted stress tests on various system operations, including queries and writes. These tests were designed to evaluate the system’s ability to handle large volumes of simultaneous requests and to ensure it met the performance criteria required for real-world applications.

The results were promising: even when the concurrent request count reached 80,000, the system maintained an average response time within 200 ms. This indicates that the system can handle significant loads without compromising on speed, ensuring a smooth and efficient user experience. Additionally, the throughput generally met business requirements, demonstrating the system’s capacity to support high transaction volumes effectively. Figure 4 illustrates the specific test results, showcasing the system’s robust performance under heavy load conditions and validating its readiness for deployment in a high-demand supply chain environment.

In addition, we also engaged a third-party security service provider to conduct penetration testing on the system, covering multiple aspects including web applications, mobile applications, blockchain nodes, etc. After multiple rounds of testing and remediation, the final assessment revealed no significant security vulnerabilities. Overall, after comprehensive functional validation, performance stress testing, security assessment, and other series of tests, all system metrics met the expected requirements [10]. The system can stably and efficiently operate in a production environment, providing secure and reliable traceability services for enterprises. In the future, we will continue to monitor the latest technological developments, strengthen performance optimization, security protection, and other efforts to continuously enhance the reliability and applicability of the system.

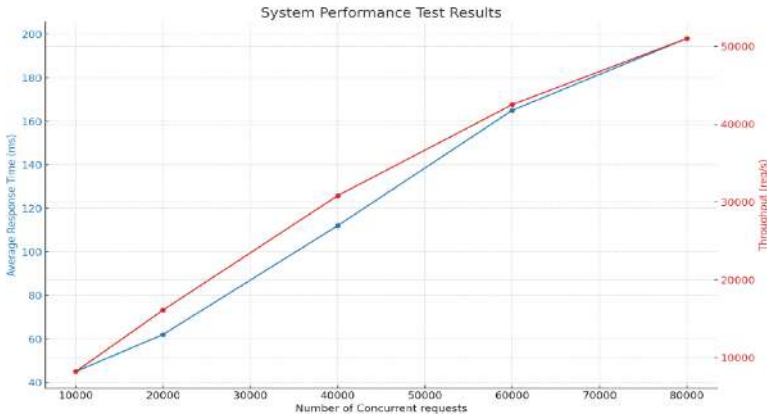


Fig. 4. System Performance Test Results

5 Conclusion

The application of blockchain technology in supply chain traceability has demonstrated immense value. By constructing a secure, trustworthy, transparent, and open data platform, enterprises can comprehensively track the entire process of products from raw material procurement to final sales, eliminate counterfeit products, maintain brand image, and win consumer trust. Additionally, the traceability system can not only improve internal operational efficiency but also promote data sharing and collaboration between upstream and downstream partners, creating a highly interconnected supply chain ecosystem. However, realizing a truly efficient and reliable blockchain traceability system requires optimization design in multiple aspects such as performance, security, scalability, and deep integration with existing enterprise IT architectures. Only through comprehensive testing, validation, and continuous optimization and iteration, can this innovative technology create maximum value for enterprises and drive the digital transformation of the entire supply chain.

Acknowledgement. This work is supported by the 2024 Annual School-level Scientific Research Project of Guangdong University of Science and Technology (Project Number: GKY-2024KYYBW-36).

References

1. Pradana, I.G.M.T., Djatna, T., Hermadi, I., et al.: Readiness assessment framework and integrated participatory development approach for blockchain-based traceability systems: case study in Kintamani coffee agroindustry supply chain. *J. Soc. Comput.* **5**(4), 344–362 (2024). <https://doi.org/10.23919/JSC.2024.0027>
2. Li, J., Wang, Z., Guan, S., et al.: ProChain: a privacy-preserving blockchain-based supply chain traceability system model. *Comput. Indust. Eng.* **187**, 109831 (2024). <https://doi.org/10.1016/j.cie.2023.109831>

3. Yuan, J., Ding, J., Ma, Y., et al.: Optimized design and implementation of testing laboratory based on blockchain technology. *J. Phys: Conf. Ser.* **2219**(1), 012037 (2022)
4. Ge, Y., Zhang, W., Li, H., Xie, Z.: Traceability analysis of distributed new energy system access to green power based on blockchain consensus algorithm. *Int. J. Emerg. Electr. Power Syst.* **25**(4), 523–532 (2024). <https://doi.org/10.1515/ijeeps-2023-0092>
5. Emroozi, V.B., Roozkhosh, P., Modares, A., et al.: A new supply chain design to solve supplier selection based on internet of things and delivery reliability. *J. Ind. Manage. Optimiz.* **19**(11), 7993–8028 (2023). <https://doi.org/10.3934/jimo.2023028>
6. Bali, S., Bali, V., Mohanty, R.P., et al.: Analysis of critical success factors for blockchain technology implementation in healthcare sector. *Benchmarking: An Int. J.* **30**(4), 1367–1399 (2023). <https://doi.org/10.1108/BIJ-07-2021-0433>
7. Sharma, A., Sharma, A., Singh, R., et al.: Blockchain adoption in agri-food supply chain management: an empirical study of the main drivers using extended UTAUT. *Bus. Process. Manag. J.* **29**, 737–756 (2023). <https://doi.org/10.1108/bpmj-10-2022-0543>
8. Nie, Z.X., Long, Y.Z., Zhang, S.L., et al.: A controllable privacy data transmission mechanism for Internet of things system based on blockchain. *Int. J. Distrib. Sens. Netw.* **18**(3), 303–315 (2022). <https://doi.org/10.1177/15501329221088450>
9. Zhao, S., Li, W.: Blockchain-based traceability system adoption decision in the dual-channel perishable goods market under different pricing policies. *Int. J. Prod. Res.* **61**, 4548–4574 (2023). <https://doi.org/10.1080/00207543.2023.2168309>
10. Tao, Z., Chao, J.: The impact of a blockchain-based food traceability system on the online purchase intention of organic agricultural products. *Innov. Food Sci. Emerg. Technol.* **92**, 103598 (2024). <https://doi.org/10.1016/j.ifset.2024.103598>



A Data Synchronization Desensitization System Based on Format-Preserving Encryption

Jingao Chen¹✉, Xiaoxu Hu², Peng Zhang¹, Yu Guo¹, and Yunjun Ma¹

¹ School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

chenjingao@njust.edu.cn

² National Computer Network Response Technical Team/Coordination Center Technology, Beijing, China

Abstract. Data desensitization is the process of performing some special transformation on sensitive data to ensure data security. Data synchronization technology is widely used in relational databases to ensure data consistency and synchronization between different databases. The current data synchronization system is unable to preserve the original format of data while desensitizing sensitive data and is unable to adopt different desensitization methods for different types of sensitive data. In order to solve the above problems, a desensitization algorithm MixFF3 that preserves the original format of the data is proposed, which combines the data mixing and FF3-1 algorithms to achieve data desensitization. A data synchronization desensitization system is designed, which uses CDC technology to capture changed data in real time and can flexibly desensitize different kinds of data with multiple methods and synchronize them to the target database. We tested the performance of the MixFF3-based synchronous desensitization system for synchronous desensitization on data tables of different magnitudes. The results show that the system preserves the original format of the data and does not impose excessive overhead on the data synchronization task after desensitization using MixFF3 for a 100,000-record synchronization task.

Keywords: Data Desensitization · Format Preserving Encryption · Data Synchronization · FF3-1 · Data Mixing

1 Introduction

In today's digital age, data synchronization systems play a vital role in various industries to ensure that data between different systems is always consistent and available [1]. However, when a data synchronization system synchronizes data from a source database to a target that requests the data, in addition to the data that is needed for analysis, there is also the possibility of exposing sensitive personal or organizational information to the target as well. Currently, most data synchronization systems do not support data desensitization, and a few synchronization systems that do support data desensitization, such as CloudCanal, only use a simple masking method to achieve data desensitization,

such as replacing every character in the sensitive data with "*" [2, 3]. Although this method is simple, it destroys the original data format, making the desensitized data unable to directly meet the use of the target end. In addition, the same desensitization method is used for all sensitive data, which lacks flexibility and fails to meet the demand for different desensitization methods for different fields.

In view of the above problem, a preserved-format encryption desensitization algorithm based on FF3-1 and data mixing, and a data synchronization desensitization system that can use different desensitization strategies for different sensitive data are proposed. The desensitization algorithm proposed in this paper mixes the data to be encrypted with the interfering data before encrypting it using the FF3-1 algorithm, which is called MixFF3. It improves the randomness and complexity of the existing standard FF3-1 algorithm while preserving the original data format. The proposed data synchronization and desensitization system can extract the data from the data source in real time, and after different desensitization methods are applied to different sensitive data in the records, the desensitized data are synchronized to the target end, which improves the flexibility and usability of data synchronization and desensitization.

2 Method

2.1 FF3-1 Algorithm

FPE (Format-Preserving Encryption) is a special type of encryption algorithm that encrypts the input data into a ciphertext with the same format as the original data [4–6], thus preserving the format of the data. FF3-1, as an improved version of FF3, is especially suited for encrypting short data and data with special formatting requirements. FF3-1 not only provides relatively good performance and security [7], but also provides efficient encryption while preserving the data format, making it an ideal choice for handling short and sensitive data such as cell phone numbers.

2.2 Data Mixing-Splitting Algorithm

The FF3-1 encryption algorithm is a deterministic encryption mode where the same plaintext yields the same ciphertext every time it is encrypted [8]. The lack of probabilistic encryption makes it possible to perform frequency or pattern analysis on the ciphertext. The limited randomness leads to an attacker being able to break the encryption algorithm by methods such as statistical analysis or frequency analysis to gain access to the original data. Vidhya proposed a Data Mixing-Splitting Algorithm that mixes the original data with the interfering data before encrypting it and splits the data after encrypting it [9]. This algorithm generates the ciphertext by crossing multiple fields to generate the plaintext, which improves the randomness and complexity of the existing standard FF3-1 algorithm.

Data Mixing-Splitting Algorithm consists of two parts which are Data Mixing Algorithm as well as Data Splitting Algorithm. The data mixing algorithm mixes the plaintext with the interfering data to generate the post-mixed data for FF3-1 encryption. The position of each character of the interfering data in the data mixing algorithm to be inserted in the plaintext is obtained by using the key of FF3 as shown in Algorithm1.

```

Data: Plaintext X1 ,Interference data X2, Key for FF3 K
Result: Mixed data X
1  X1_digits ← list(map(int, X1));
2  X2_digits ← list(map(int, X2));
3  X1_len ← len(X1_digits);
4  X2_len ← len(X2_digits);
5  X_len ← X1_len + X2_len;
6  X ← [0] * (X1_len + X2_len);
7  keyset ← set();
8  foreach digit in K do
9      keyset.add(int(digit, 16));
10 end
11 keylist ← sorted(list(keyset));
12 positions ← [(element * keylist[-1]) % X_len for element in keylist];
13 positions ← positions[:X2_len];
14 index ← 0;
15 foreach i in X2_digits do
16     X[(positions[index])] ← i;
17     index ← index + 1;
18 end
19 i ← 0;
20 for index in range(0,X_len) do
21     if index not in positions then
22         X[index] ← X1_digits[i];
23         i ← i+1;
24     end
25 end
26 X ← "".join(str(x) for x in X);

```

The data splitting algorithm splits the FF3-1 encrypted data into the encrypted plaintext as well as the encrypted jammed data and discards the latter. The data splitting algorithm obtains the positions of the interfering data in the same way and discards these characters.

2.3 MixFF3 Algorithm

In Vidhya's data mixing algorithm, each character in the FF3 key is multiplied by its maximum value, and then the product of the two is modeled on the sum of the lengths of the plaintext and the interfering data to form the elements of a new list, and take some of the elements of this list to form a position list. This method suffers from the problem that when the length of the plaintext data is small, the elements in the position list are duplicated. For this case, we have improved the computation of the position list in the data mixing algorithm to make it applicable to plaintext data of shorter length, as shown in Algorithm 2.

```

1 tmp ← -1;
2 while True do
3     positions ← [(element * keylist[tmp]) % X_len for element in keylist];
4     if len(set(positions[:X2_len])) == X2_len then
5         break;
6     end
7     if tmp + len(keylist) == 0 then
8         break;
9     end
10    tmp ← tmp + 1 ;
11 end

```

A MixFF3 desensitization algorithm for encrypting a portion of plaintext data is proposed based on the data mixing algorithm and FF3-1 algorithm after improving the positional computation method.

MixFF3 divides the plaintext data into front, middle, and back. Front and back generate the interfering data for data mixing and the tweak for the FF3-1 algorithm. The middle part serves as the plaintext data for data mixing and encryption. The lengths of the front, middle and back sections can be defined by the user.

Data: Plaintext plaintext
Result: Ciphertext ciphertext

```

1 key ← secrets.token_bytes(16).hex();
2 alphabet ← "0123456789";
3 n1,n2 ← user-defined,user-defined;
4 front ← plaintext[:n1];
5 back ← plaintext[n1 + n2:];
6 middle ← plaintext[n1 + n2];
7 Interference ← front+back;
8 hash_object ← hashlib.sha256();
9 hash_object.update(Interference.encode('utf-8'));
10 tweak ← hash_object.digest()[:7].hex();
11 c ← FF3Cipher.withCustomAlphabet(key, tweak, alphabet);
12 mixdata ← data_mixing(middle,Interference,key);
13 cipher_mixdata ← c.encrypt(mixdata);
14 cipher_middle ← data_splitting(ciphertext,Interference,key);
15 ciphertext ← front + ciphermiddle + back;

```

MixFF3 mixes the middle part with the interfering data and then encrypts it using the FF3-1 algorithm and then uses the data splitting algorithm to get the encrypted middle data. The encrypted middle data is spliced with front and back to be used as the desensitized data as shown in Algorithm 3.

3 Design of System

3.1 Architecture of Data Synchronization Desensitization System

Data synchronization and desensitization system can monitor and capture the source database changes in real time and desensitize the sensitive fields in the changed data and insert them into the target database to achieve real-time desensitization and synchronization of data. It mainly includes Change Data Capture module and Data Desensitization module. The overall framework of the system is shown in Fig. 1.

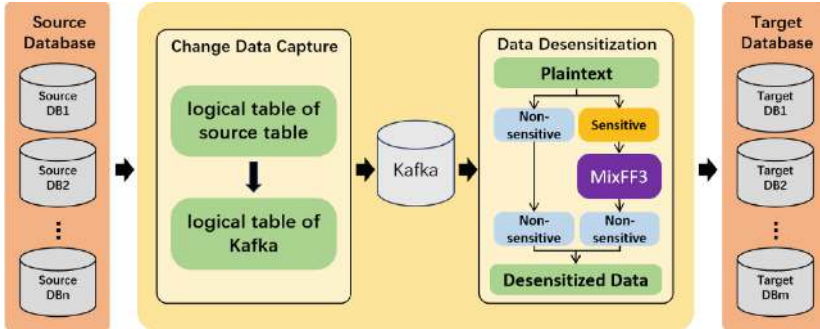


Fig. 1. System Framework.

Change Data Capture module uses CDC technology [10] to monitor and capture database changes (including data insertion INSERT, update UPDATE, deletion DELETE, etc.) in real time, record these changes in the order in which they occur, and quickly write them to Kafka.

The Data Desensitization module desensitizes the sensitive parts of the data in the Kafka topic and inserts them into the target database to achieve real-time synchronization and desensitization of data.

3.2 Change Data Capture

The Change Data Capture module uses FlinkCDC to capture change data and requires the source database to have CDC-related features enabled. Take MySQL as an example, you need to enable the Binary Log function of MySQL to capture the change events in the database.

In each synchronization desensitization task, the system creates the logical table of the source table and the logical table of Kafka in Flink to realize the monitoring of the source table and Kafka topics. Change data capture is achieved by extracting the change events of the source table into Flink in real time, formatting them and writing them to Kafka.

Definition Of Logical Table. The definition of the logical table includes the structure of the table (fields and data types) and the configuration of the connection to the actual data source. The logical table of the source table and the logical table of Kafka are basically the same in the definition of the table structure. When defining the structure of the logical table, you need to map the data types of each field in the source database to FlinkSQL types according to the Type Mapping Table. The main aspects of connection configuration are the type of FlinkCDC connector used, connection information for the source database and Kafka, etc.

Type Mapping Table. Flink cannot directly use the data types in the source data table when creating a logical table, it needs to be mapped to FlinkSQL types. The data type mapping of some common databases to Flink is shown in Table 1.

Table 1. Type Mapping Table.

MySQL	Oracle	Hive	Flink SQL
boolean	NUMBER(1)	boolean	BOOLEAN
int	NUMBER(10,0)	int	INT
double	BINARY_DOUBLE	double	DOUBLE
varchar	VARCHAR2	string	STRING
char	CHAR	string	STRING
text	CLOB	string	STRING
date	DATE	---	DATE
datetime	---	timestamp	TIMESTAMP
timestamp	TIMESTAMP	timestamp	TIMESTAMP
decimal	NUMBER	---	DECIMAL

3.3 Data Desensitization

The data processing module consumes messages in Kafka topics in real time and examines the message’s operation type to perform different processing logic, as shown in Fig. 2.

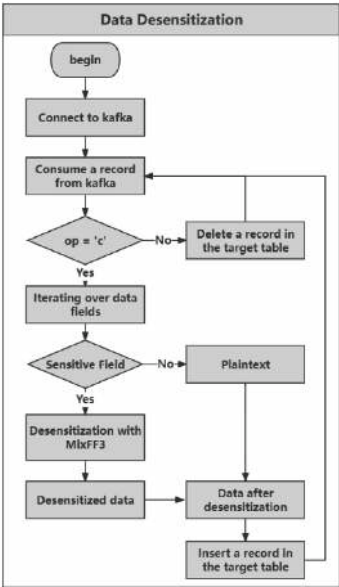


Fig. 2. Process of Data Desensitization.

If the action type of the message is ‘c’, it means that this message is inserting a piece of data into the target database. Get the specific database record in the message and iterate over the fields in it. If the field name is in the Field Desensitization Config Dictionary, take the value of the field as a parameter, call the corresponding desensitized method, and update the desensitized value to the data to be inserted. After the desensitization process is completed, the desensitized data is inserted into the target database. If the operation type of the message is ‘d’, it means this message is to delete a data from the target database. Get the primary key of the specific database record in the message and delete the data from the target database according to the primary key.

The only operation types of changed data captured by CDC are insertion operation ‘c’ and deletion operation ‘d’, and the update operation is split into deletion and insertion operations in CDC.

Field Desensitization Config Dictionary Configured by the user, marks which fields in the source data table are sensitive and the desensitization method used for the sensitive fields. The key of the dictionary is the field name of the source data table and the value is the method name in the Desensitization Methods Table. The format of the dictionary is shown in Table 2.

Table 2. Field Method Mapping.

Sensitive Field	Desensitization Method
Phone Number	Phone_MixFF3
ID Number	ID_MixFF3
Credit Card Number	Credit_Card_MixFF3

Desensitization Methods Table. The system maintains a desensitization method table in the backend, and the two main fields of the table are desensitization method name and desensitization method string. Each desensitization method name corresponds to a unique desensitization method, and the specific implementation of the method is recorded in the desensitization method string.

The system has predefined desensitization methods for some sensitive data types, such as Phone_MixFF3 for cell phone numbers and ID_MixFF3 for ID card numbers. These predefined desensitization methods are based on MixFF3, only the length of the front, middle and back parts are different. In addition to the predefined MixFF3-based desensitization methods, users can also customize non-MixFF3-based desensitization methods.

When a synchronized desensitization task is created, the Desensitization Methods Table is queried based on the set of values in the Field Desensitization Config Dictionary, and the desensitization method string is written to the script of the data processing process.

4 Experiment

4.1 Experiment Setting

The experimental server has an Intel Xeon Platinum 8369B CPU, 48 GB of total memory, and an operating system of Alibaba Cloud Linux 3. The version of Flink used for the data synchronization desensitization system is 1.13, and the version of Kafka is 2.2.1. The versions of both the source and target databases used for testing are MySQL 5.7.42.

4.2 Desensitization Example

Some of the raw sensitive data from the tests are shown in Table 3:

Table 3. Plaintext.

Phone Number	Credit Card Number
13792774562	6228484567890123
13392774592	6228483456789012

For a Phone Number of length 11, MixFF3 sets the length of front, middle and back to 3, 4 and 4 respectively, and encrypts only the part from the 4th to the 7th digit. For Credit Card Number of length 16, the length of front, middle and back in MixFF3 is set to 4, 8 and 4 respectively, and only the part from 5th to 12th digit is encrypted.

The data after desensitization are shown in Table 4:

Table 4. Desensitized Data.

Phone Number	Credit Card Number
13761544562	6228426459200123
13359374592	6228919964709012

4.3 Performance Analytics

Data synchronization desensitization is performed on 10000, 50000 and 100000 records respectively, each record has only one sensitive field and the results of the experiment are shown in Table 5:

Table 5. Performance of Different Desensitization Methods.

Method	# of Records	Exec.Time(sec)	Encryption.Time(sec)
Clear text	10000	209.92	- - -
Clear text	50000	1136.96	- - -
Clear text	100000	2301.80	- - -
FF3-1	10000	202.48	1.13
FF3-1	50000	1154.42	5.15
FF3-1	100000	2332.30	11.84
MixFF3	10000	216.62	1.62
MixFF3	50000	1194.27	7.23
MixFF3	100000	2347.51	14.23

The synchronization time to the target database without desensitization for 10000, 50000, and 100000 records is 209.92, 1136.96, and 2301.80 s respectively. The synchronization and desensitization time without data mixing and only using FF3-1 for encryption are 202.48 s, 1154.42 s, and 2332.30 s, respectively. The time taken for FF3-1 encryption was 1.13 s, 5.15 s, and 11.84 s, respectively. The time taken for synchronized desensitization using MixFF3 is 216.62 s, 1194.27 s, and 2347.51 s. The time taken for MixFF3 encryption was 1.62, 7.23, and 14.23 s, respectively.

The desensitization process inevitably increases the data synchronization time, but the increase is relatively small and within acceptable limits. The overhead of MixFF3 encryption is only slightly higher than that of FF3-1, but provides additional security. For 100,000 records, MixFF3 encryption took only 2.39 s longer than FF3-1 encryption, indicating that its additional overhead is not significant in the case of large data volumes.

5 Conclusion

A reserved format encryption desensitization algorithm based on FF3-1 and data mixing is proposed to mix the data to be encrypted with the interfering data before encryption using the FF3-1 algorithm, which improves the randomness and complexity of the existing standard FF3-1 algorithm. A data synchronization desensitization system that can use different desensitization methods for different sensitive data is designed, which can flexibly use different desensitization methods for different sensitive data and synchronize them to the target database in real time. The results show that this system does not cause excessive overhead to the data synchronization task when realizing data desensitization, preserves the original format of the data and has better security.

References

1. Györödi, C.A., Turtureanu, T., Györödi, R.Ş, Zmaranda, D.R.: Implementing a synchronization method between a relational and a non-relational database. *Big Data Cognit. Comput.* **7**(3), 153 (2023). <https://doi.org/10.3390/bdcc7030153>
2. Goyal, C.: Data masking: need, techniques & solutions. *Int. Res. J. Manag. Sci. Technol. (IRJMST)* **6**(5), 221–229 (2015)
3. Kumargk, R., Justus Rabi, B., Manjunath, T.N.: A Study on dynamic data masking with its trends and implications. *Int. J. Compu. Appl.* **38**(6), 19–24 (2012). <https://doi.org/10.5120/4612-6828>
4. Wang, H., Zhang, Y., Guo, Z., Li, T., Fanl, L.: Format preserving encryption of sensitive data in database. 12718, 127180T (2023). <https://doi.org/10.1117/12.2681546>
5. Li, S., Gao, Y., Shi, T.: Identity-based hybrid format-preserving encryption scheme. In: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 470–474 (2020). <https://doi.org/10.1109/ITOEC49072.2020.9141834>
6. Jang, W., Lee, S.-Y.: A format-preserving encryption FF1, FF3-1 using lightweight block ciphers LEA and, SPECK. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing, in SAC'20. New York, NY, USA: Association for Computing Machinery, pp. 369–375 (2020). <https://doi.org/10.1145/3341105.3373953>
7. Betül Durak, F., Vaudenay, S.: Breaking the FF3 format-preserving encryption standard over small domains. In: Katz, J., Shacham, H. (eds.) *Advances in Cryptology – CRYPTO 2017: 37th Annual International Cryptology Conference*, Santa Barbara, CA, USA, August 20–24, 2017, Proceedings, Part II, pp. 679–707. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-63715-0_23
8. Dunkelman, O., Kumar, A., Lambooi, E., Sanadhya, S.K.: Cryptanalysis of feistel-based format-preserving encryption (2020). Accessed: 16 Jun 2024. <https://eprint.iacr.org/2020/1311>

9. V. S, “Secure Format Preserving Encryption for Multiple Data Fields,” presented at the Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security, IACIDS 2023, 23–25 Nov 2023, Lavasa, Pune, India, Mar. 2024. Accessed 16 Jun 2024. <https://doi.org/10.4108/eai.23-11-2023.2343192>
10. Zhang, C., Qi, D., Li, W., Huang, W., Wang, X.: SimpleSync: A parallel delta synchronization method based on Flink. *Concurrency Comput.: Pract. Experience* **33**(20), e6327 (2021). <https://doi.org/10.1002/cpe.6327>



Research on the Development Strategy of Smart Tourism Driven by Big Data Platform

Yuehua Bai and Yu Yang^(✉)

School of Management, Wuhan Donghu University, Wuhan, China
yy123_dhxy@163.com

Abstract. This article explores the promotion and application of big data platforms in the development of smart tourism. Firstly, it outlines the concept and development trend of smart tourism, emphasizing the problems of information silos and single experiences in traditional tourism industry. Subsequently, the role and significance of big data platforms in smart tourism were elaborated in detail, and the driving role of big data platforms in the development of smart tourism was explored. Based on this, the potential role of big data platforms in optimizing tourism resource allocation, improving service quality, and promoting industry upgrading was further analyzed. Finally, the significance and future development trends of smart tourism driven by big data platforms are summarized, and it is emphasized that big data platforms will become the core driving force for smart tourism. Through the application of big data platforms, smart tourism can better meet the needs of tourists, promote sustainable development of the tourism industry, improve tourism experience, and promote innovative development of the tourism industry.

Keywords: Big Data Platform · Smart Tourism · Data Analysis · Personalized Recommendations · Tourism Industry Development

1 Introduction

With the development and popularization of information technology such as big data platforms, the tourism industry has entered a new era of digitization, informatization, and intelligence. Big data platforms have become an important tool to accelerate the intelligent development of the tourism industry. Traditional tourism cannot meet the needs of tourists, and there are problems such as information asymmetry and single experience. To address these issues and promote the development of big data platforms in fields such as data mining, big data analysis, and data visualization, this article aims to study the application and driving role of big data platforms in the field of smart tourism, providing theoretical basis and practical reference for the development strategy of smart tourism [1]. To better introduce the role and role of big data platforms in smart tourism, the article first explains the concept of smart tourism and the problems existing in traditional tourism industry. Then, the article explores in detail the role and role of big data platforms in smart tourism, including data collection, storage, analysis, and application.

It analyzes the driving role and advantages of big data platforms in the development of smart tourism, explores the application fields of big data platforms in smart tourism, and focuses on the application cases and values of personalized recommendations, real-time intelligent decision-making, and tourist behavior prediction. On this basis, this article further analyzes the potential role of big data platforms in improving the quality of tourism services and optimizing resource allocation. Finally, the significance and future development trends of smart tourism driven by big data platforms are summarized, emphasizing the important role of big data platforms in promoting the transformation and upgrading of the tourism industry and advancing the development of the digital economy.

2 Development of Smart Tourism Driven by Big Data Platforms

2.1 Overview of Big Data Platforms and Common Algorithms

Big data platforms usually include functional modules such as data collection, data storage, data processing, and data analysis. The purpose of data collection is to obtain data from different data sources. In smart tourism, various types of information can be obtained through sensors, mobile devices, social media, and other channels, such as user behavior, location information, traffic conditions, climate data, etc. The diversity of data provides a rich data foundation for smart tourism. Data storage is the process of storing collected data in a structured or unstructured form in a database or data warehouse for subsequent processing and analysis [2]. The commonly used data storage technologies in big data platforms include distributed file systems, NoSQL databases, and data warehouses. Data processing is the core link of big data platforms, mainly cleaning, transforming, and integrating collected and stored data. Big data platforms often use distributed computing frameworks to process large-scale and high-speed data, preparing for subsequent analysis and application. These technologies can mine user preferences, predict user behavior, optimize tourism resource allocation, and promote the development of smart tourism.

2.2 Current Status and Challenges of Smart Tourism Development

With the continuous development of technologies such as big data, cloud computing, and artificial intelligence, smart tourism is rapidly popularizing and promoting. Smart tourism is a comprehensive intelligent upgrade of the tourism industry using advanced information technology to improve the quality of tourism services, industry competitiveness, and socio-economic benefits. The emergence of new technologies will inevitably bring many challenges, and the development of smart tourism is no exception. Information asymmetry is a core issue in development. The information obtained by tourists is usually single and one-sided, which is difficult to accurately reflect the true quality and value of tourism products, resulting in many uncertainties in the selection and consumption process. Due to the inconsistent implementation and standards of smart tourism technology in different regions and enterprises, the standardization of tourism services is limited. This also limits the overall effectiveness of smart tourism development and the improvement of user experience. With the centralized storage and sharing

of smart tourism data, the risk of abuse and leakage of personal and payment information of tourists has become another important challenge for the development of smart tourism. On the basis of addressing these challenges, promoting the development of smart tourism can bring greater changes and development opportunities to the tourism industry. With the increase in the number of tourists, the service needs and types of the tourism industry have become increasingly complex and diverse, making it impossible to ensure the standardization, standardization, and stability of services. The quality and safety of tourism products are related to the long-term development and brand image of the tourism industry. In the traditional tourism industry, it is difficult to conduct comprehensive and multi-dimensional quality supervision and management of tourism products, resulting in uneven quality of tourism products. There are also problems such as imbalanced allocation of tourism resources, fierce industry competition, and outdated facilities and equipment, all of which constrain the development and promotion of smart tourism. In response to these challenges and issues, big data platforms provide new solutions for the development of smart tourism [3].

2.3 Role and Advantages of Big Data Platforms in Smart Tourism Development

With the popularization of intelligent devices and the Internet, tourists' behavior, location, preferences and other types of data are constantly generated. Big data platforms can acquire rich tourism data through multiple channels of data collection and integration, play an important role in data collection, and form a comprehensive data resource database. The traditional tourism industry often faces the challenge of massive data storage and management. By utilizing distributed file systems and cloud platform technology, big data platforms have high scalability and fault tolerance, which can store and manage large-scale tourism data. Data storage and management have advantages. Customized data management solutions at the same time facilitate the improvement of data utilization efficiency and security. In addition, big data platforms also play an important role in data processing and analysis. With the help of distributed computing technology and machine learning algorithms, they can process and analyze tourism data in real-time and quickly. They can easily extract valuable information and patterns from a large amount of tourism data, providing data support and guidance for decision-making in smart tourism. Not only that, he can also achieve personalized recommendations for tourism information [4]. By analyzing user behavior data, big data platforms can accurately understand user preferences and needs, predict tourist behavior, provide personalized tourism recommendation services, improve tourist satisfaction, predict tourist behavior, and make reasonable resource allocation and service arrangements to improve the efficiency of tourism resource utilization. These advantages make big data platforms an important driving force for the development of smart tourism, promoting innovation and upgrading of the tourism industry.

3 Application of Big Data Platform Algorithms in Smart Tourism

3.1 Application of Data Mining and Machine Learning Algorithms in Tourist Behavior Analysis

The role of data mining is to automatically detect, extract, and recognize hidden patterns and correlations from massive data, while machine learning is different. It belongs to a technology that allows machine systems to continuously optimize themselves by learning and analyzing data. In the field of tourism, using data mining algorithms can provide in-depth insights into the interests, consumption habits, and preferences of tourists for tourist attractions, dining, and shopping. Through analyzing the results, smart tourism platforms can create personalized travel itineraries, restaurant choices, and shopping recommendations for tourists, which can significantly improve their satisfaction[5]. Machine learning mainly uses rich historical data to predict tourist behavior patterns and establish predictive models for purchasing behavior. Machine learning algorithms can predict future consumption trends and purchase intentions. In this way, smart tourism platforms can prepare corresponding tourism resources and services in advance, ensure efficient utilization of resources, and meet the immediate needs of tourists. Figure 1 shows the process of the K-means algorithm. Based on the location information and historical trajectory of LBS, popular attractions and tourist routes can also be discovered, helping smart tourism platforms optimize resources and manage traffic. It can also analyze the emotional evaluation and comments of tourists to understand their satisfaction and dissatisfaction with services, and improve and optimize services. The application of data mining and machine learning algorithms provides strong support for the development of smart tourism and accelerates the promotion and popularization of smart tourism.

3.2 Application of User Interest Model and Behavior Prediction Algorithm in Personalized Tourism Recommendation

Big data platforms play a crucial role in the field of smart tourism. By constructing refined user interest models and adopting advanced behavior prediction algorithms, they provide customized and personalized tourism recommendation services to tourists, recommending the most suitable tourist resources such as attractions and restaurants to meet their needs, allowing them to enjoy a more intimate and enjoyable experience on the journey [6]. A personalized travel recommendation system can more accurately understand user needs by analyzing their interest models, and provide personalized travel recommendations based on their interests. The architecture of the personalized travel recommendation system is shown in Fig. 2.

The application of behavior prediction algorithms has also brought great convenience to tourists. By analyzing users' historical travel behavior data and personal preferences, big data platforms can predict their future travel trends and preferences, recommend the most suitable travel destinations for them, and help them develop personalized travel plans. In addition to the predictive ability mentioned above, it can also help tourists better plan their itinerary, recommend relevant attractions, hotels, and transportation methods, making the entire travel process smoother and more comfortable. In summary, big data

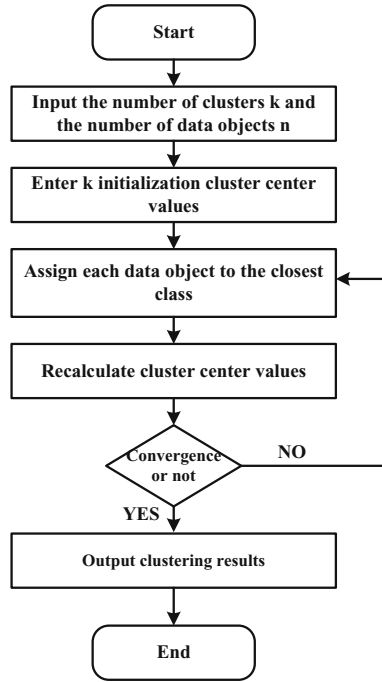


Fig. 1. The process of K-means algorithm.

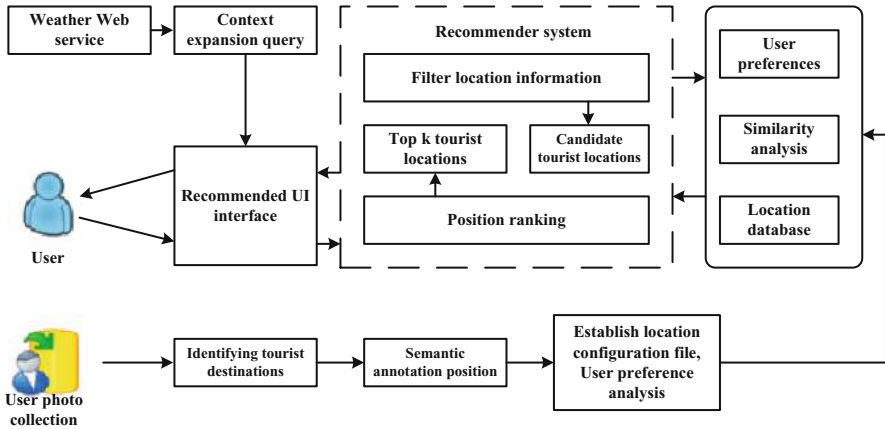


Fig. 2. Architecture of a personalized tourism recommendation system.

platforms provide tourists with precise and personalized travel recommendation services by constructing user interest models and utilizing behavior prediction algorithms. This not only improves user satisfaction and experience, but also effectively promotes the development and popularization of smart tourism [7].

3.3 Application of Prediction and Optimization Algorithms in Tourism Market Analysis and Marketing Decision-Making

Big data platforms can assist tourism industry players in market analysis, optimizing resource allocation, and making more accurate and effective marketing decisions through the application of prediction and optimization algorithms. They can analyze historical tourism data, market trends, and related economic indicators. Prediction algorithms can be used in tourism market analysis to predict future tourism demand for a certain period of time. Based on the prediction results provided by the platform, customers should make reasonable adjustments to tourism, prepare resource allocation in advance, and optimize business plans. Optimization algorithms can also utilize mathematical models and computational techniques to assist tourism operators in formulating optimal pricing strategies, product combinations, and sales channel arrangements [8]. They can search for the best decision-making solutions and strategies in tourism marketing decisions to optimize resource utilization, improve profitability, optimize recommendation systems, analyze user behavior data, provide more attractive personalized recommendations, thereby improving user satisfaction and loyalty, achieving more efficient resource management and service arrangements, enhancing the efficiency of tourism resource utilization, and playing a positive role in tourism safety and crisis management.

4 Smart Tourism Development Strategies Driven by Big Data Platforms

4.1 Innovative Smart Tourism Models

Driven by big data platforms, the innovation of smart tourism models has brought many new opportunities and challenges to the tourism industry. The innovative smart tourism model provides a more personalized and customized tourism experience. It analyzes user historical behavior data and user profiles to customize tourism routes based on user interests, and provides personalized scenic spot recommendations to improve user travel experience and satisfaction. The smart tourism model can also integrate resources from all parties, achieve optimization and collaboration of the tourism supply chain. The smart tourism platform can use cloud technology and big data analysis to understand the real-time flow of people and resource utilization of each scenic spot, optimize tourist flow and resource allocation, and improve resource utilization efficiency [9]. The innovative smart tourism model also focuses on multi-channel and all media marketing strategies. With the support of big data platforms, smart tourism platforms can integrate online and offline marketing channels, such as publishing real-time travel information and discount information through social media platforms, providing online booking and navigation functions through mobile applications, and providing real-time navigation and recommendation services through smart devices, thereby attracting more tourists and enhancing consumer loyalty. The innovative smart tourism model, driven by big data platforms, brings new opportunities and prospects for the development of smart tourism. The personalized smart tourism model can meet the needs of different tourists and enhance the efficiency and competitiveness of the tourism industry.

4.2 Tourism Market Analysis and Forecasting Strategies

Businesses can efficiently integrate and analyze massive amounts of tourism data using big data platforms, including user behavior, traffic conditions, scenic spot traffic, etc., in order to analyze the scale, structure, and development trends of the tourism market. By utilizing these analysis results, not only can consumer behavior and needs be revealed, but also prediction algorithms and models can be used to predict future market development trends for tourism operators. Big data platforms can not only support market analysis and prediction of smart tourism, but also help tourism operators collect user feedback and adjust products and services in a timely manner through online promotion. Therefore, tourism operators can develop more competitive strategies based on accurate market data analysis and prediction, and promote the innovative development of smart tourism [10].

4.3 Smart Tourism Experience and Service Optimization Strategies

The core of smart tourism strategy lies in personalized and customized services. Merchants utilize big data platforms and advanced algorithms to achieve intelligent recommendations and customized experiences, optimize service quality and efficiency, and thereby enhance passenger satisfaction and loyalty. By using precise recommendation algorithms and location services, personalized recommendations such as scenic spots and catering are tailored for each traveler to meet different needs and enhance the overall travel experience. At the same time, big data analysis can grasp passenger needs and feedback in real-time, flexibly adjust service processes, optimize resource allocation, reduce waiting time, and achieve more efficient and considerate services. The strategy for optimizing the experience and services of smart tourism focuses on integrating various intelligent technologies and innovative means, providing more intelligent and convenient tourism services [11]. By applying technologies such as artificial intelligence and the Internet of Things, smart tourism platforms can provide innovative service methods such as voice guidance, intelligent navigation, and virtual reality, enabling travelers to more conveniently access information, navigation, and visit, enhancing the convenience and fun of the tourism experience. Tourists can use personalized and customized tourism services, as well as integrate intelligent technologies and innovative means, to provide users with more intelligent and personalized tourism experiences, meet the diverse needs of travelers, and thus enhance the competitiveness and profitability of the tourism industry.

4.4 Tourism Resource Development and Management Strategies

By applying the technology and algorithms of big data platforms, tourism operators can develop and manage tourism resources more scientifically and effectively, achieve optimal utilization of resources, improve the quality and competitiveness of tourism. Tourism resource development strategies are the foundation of smart tourism development. Through the analysis function of big data platforms, tourism operators can deeply understand the types, distribution, and value of tourism resources, quickly identify potential tourism resources, including natural landscapes, cultural heritage, distinctive customs, etc. By analyzing user search and browsing behavior, they can discover

user interests in certain specific scenic spots and activities, and then develop and promote related tourism products, enhancing the economic and social value of tourism resources. The data collection and analysis capabilities of big data platforms provide strong support for tourism resource management. By analyzing tourist behavior data, scenic area traffic data, and environmental protection data, tourism operators can understand the utilization status and environmental impact of tourism resources in real time, adjust and optimize resource development and management strategies in a timely manner, monitor the number and flow trajectory of tourists in real time, control the entrance flow of scenic areas, guide tourist routes reasonably, and reduce excessive consumption of resources and environmental pollution. The development and management strategies of tourism resources also need to focus on cooperation and sharing. Through the information sharing and collaborative cooperation functions of big data platforms, tourism operators can cooperate with other related industries to achieve the sharing and optimization of tourism resources. They can cooperate with transportation departments to incorporate traffic flow and travel demand data into big data platforms, optimize the allocation of transportation resources through analysis and prediction, and improve the convenience and efficiency of tourism. Through scientific resource development, meticulous resource management, and cooperative sharing strategies, the economic and social benefits of tourism resources can be effectively improved, achieving sustainable development of the tourism industry [12].

4.5 Smart Tourism Marketing and Promotion Strategies

Tourists can accurately grasp market demand through the application of big data platform technology and algorithms, and adopt targeted marketing and promotion strategies to improve brand awareness, attract tourists, and increase sales. Smart tourism marketing strategies can accurately target user groups and achieve personalized marketing. Merchants use big data technology to collect user data, deeply analyze user behavior, interests, and consumption habits, and achieve user segmentation and personalized label setting. Based on these precise information, merchants can provide customized marketing activities and tourism services to users, meet their personalized tourism needs, increase their dependence and favorability towards tourism brands, and thereby enhance their purchase intention and loyalty. The promotion strategy of smart tourism cannot be separated from the network promotion and market analysis functions of big data. These functions help businesses attract target users through multi-channel promotion, and deeply understand market characteristics, competitive trends, and potential needs, providing strong support for precision marketing. Combining innovative technological means such as virtual reality and augmented reality to provide a unique tourism experience, attract more user attention and participation, the implementation of smart tourism marketing and promotion strategies is of great significance for the successful implementation and promotion of smart tourism [13].

5 Smart Tourism Cases Driven by Big Data Platforms

5.1 Construction and Application Case of Smart Tourism Platform

Tourists can utilize the technology and algorithms of big data platforms to develop smart tourism platforms, achieving centralized management, intelligent analysis, and personalized recommendation of tourism information, in order to improve the tourism experience and service quality. The architecture of the smart tourism platform is shown in Fig. 3.

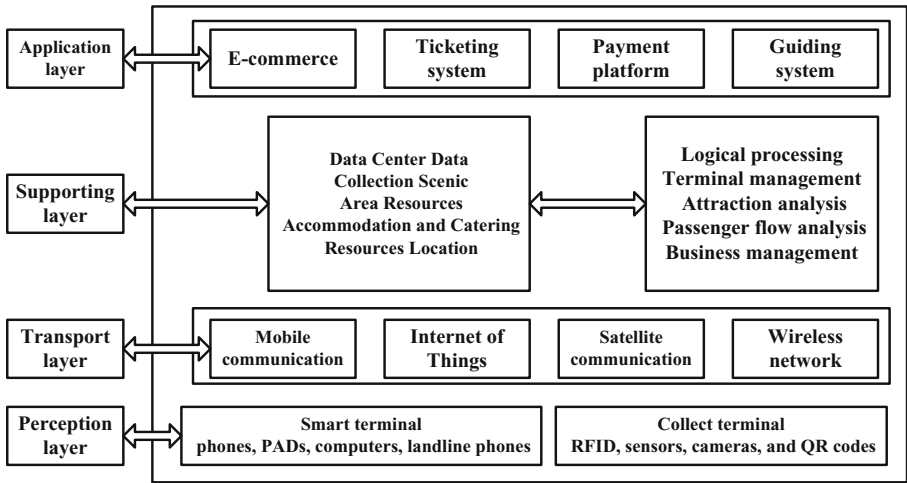


Fig. 3. Intelligent tourism platform architecture.

Taking the construction and application of a smart tourism platform in a certain province as a case study, the province has abundant tourism resources and attracts a large number of tourists. However, the traditional tourism service model has problems such as information asymmetry, resource dispersion, and service consistency. In order to solve these problems, the province has decided to build a smart tourism platform. The construction content of the smart tourism platform includes data collection and centralized management, real-time collection of tourism related data through various data sources, storage on the big data platform, and data sharing and integration with scenic spots, hotels, and transportation departments. The platform also applies big data analysis and intelligent recommendation algorithms to analyze tourist behavior and needs, providing them with personalized scenic spot recommendations, catering recommendations, etc., improving tourist satisfaction and experience. The platform applies technology such as location services and virtual tour guides to provide tourists with real-time positioning navigation and virtual tour experience, enriching their travel experience. This case fully demonstrates the role of big data platforms in the development of smart tourism, as well as the importance of smart tourism platforms [14].

5.2 Case Study on Smart Tourism Scenic Area Management

Tourism scenic area managers can scientifically and efficiently manage scenic area resources by applying the technology and algorithms of big data platforms, improving the operational efficiency and tourist experience of the scenic area. An analysis was conducted on the practice of a smart tourism management platform in a famous scenic spot. The famous scenic spot faces problems such as a large number of tourists and difficult management. In order to improve management level and service quality, the scenic spot has built a smart tourism management platform to collect a large amount of operational data, including information on the number of tourists, flow trajectory, and satisfaction survey. Through in-depth analysis of scenic area data, managers can accurately identify hot areas, congested periods, and tourist preferences, thereby optimizing operational strategies and service arrangements, improving service quality and tourist satisfaction. At the same time, with the help of intelligent recommendation algorithms, combined with the personalized needs, historical travel behavior, and consumption habits of tourists, customized tourist attractions, dining, and shopping recommendations are provided to further enhance the tourist experience. In addition, sensors and monitoring devices installed in the scenic area track tourist flow and traffic conditions in real time, helping managers efficiently adjust pedestrian guidance and traffic organization, reduce congestion, ensure safety, optimize resource allocation, and improve the operational efficiency of the scenic area. The innovative smart tourism platform of scenic spots not only cooperates with relevant tourism suppliers and third-party service providers to achieve integrated management of tourism services, but also facilitates users to book scenic spot tickets, guide services, hotel accommodation, etc., enjoying convenient one-stop services. The application of big data platform technology and algorithms can improve the operational efficiency and tourist experience of scenic spots, and contribute to the development of the smart tourism industry.

6 Conclusion

Smart tourism has emerged with the innovation and widespread application of big data platforms. Analyzing its development trend, big data platforms play a key role in smart tourism, not only providing rich and accurate information resources, but also using data analysis and mining techniques to accurately grasp market demand and user preferences, providing strong data support for the decision-making and development of smart tourism. In the development strategy of smart tourism, it is possible to implement personalized marketing strategies and innovate smart tourism promotion strategies, achieve precise positioning and market research through data analysis and recommendation algorithms, improve user satisfaction and loyalty, and promote effectiveness and returns. At the same time, scenic area management is gradually undergoing refined operations, and the real-time monitoring and prediction functions of the platform can improve the operational efficiency and tourist experience of the scenic area. Smart tourism driven by big data platforms will further optimize user experience and improve service quality. However, attention should also be paid to potential security and privacy issues that may arise during the development of smart tourism. It is necessary to strengthen the formulation and implementation of relevant regulations to ensure the sustainable development of

smart tourism. With the technology and algorithms of big data platforms, smart tourism operators can achieve personalized marketing strategies, innovative promotion methods, and refined scenic area management, improving user experience and service quality. The development of smart tourism driven by big data platforms has broad development prospects and challenges, providing important opportunities for the transformation and upgrading of the tourism industry.

References

1. Liu, L.: Knowledge graph and hotspot topic analysis in smart tourism research. *Smart Cities* **9**(11), 77–80 (2023)
2. Ding, L.: Research on information security issues in smart tourism development in the era of big data. *Digit. Commun. World* **11**(5), 166–168 (2023)
3. Wang, K.: Discussion on the problems and countermeasures of smart tourism construction based on information technology: taking the smart tourism system project of Jianchuan Ancient Town as an example. *Enterp. Reform Manage.* **21**(3), 71–73 (2023)
4. Liu, Y., Chen, X.: Research on the development model of smart rural tourism under the background of big data. *Shanxi Agricult. Economy* **19**(6), 62–64 (2023)
5. Sheng, F., Zhang, Y., Hu, M.: Optimization path of information collaboration for smart tourism multi-subjects in the context of rural revitalization. *Econ. Res. Guide* **18**(11), 22–25 (2023)
6. Yang, C.: Research on the analysis path of smart tourism ASEB under the background of immersive experience: a case study of Mengjin District, Luoyang City, Henan Province. *Western Tourism* **18**(9), 42–44 (2023)
7. Yan, Q., Huang, X., Lin, Z.: Analysis of big data application in smart tourism. *J. Xi'an Aeronaut. Univ.* **41**(05), 67–75 (2023)
8. Han, N.: Analysis of marketing strategies of smart tourism platforms based on user experience. *Market. Front.* **16**(7), 20–22 (2023)
9. Zhao, J., Zhao, G.: Construction and research of smart tourism digital information service platform. *Commun. Inform. Technol.* **11**(4), 20–23 (2023)
10. Huang, X., Sun, X., Yuan, Y.: Discussion on the construction of smart scenic area big data management platform. *Guangdong Commun. Technol.* **43**(7), 57–63 (2023)
11. Pei, F., Wang, Y., Zhao, L.: Research on E-smart tourism supply chain with big data technology. *China New Technol. New Products* **10**(4), 32–34 (2023)
12. Zhang, X., Nan, Y., Gao, Y.: Development strategy of tourism platform in Yiliang County, Yunnan Province based on smart tourism. *Modern Horticul.* **46**(11), 84–86 (2023)
13. Zhang, L.: Research on the management mode path of smart tourism based on big data background. *Western Tourism* **24**(12), 103–105 (2022)
14. Tu, D., Gao, Z., Tang, S.: Research on information management in the development of smart tourism in the era of big data. *Jiangsu Sci. Technol. Inform.* **39**(34), 46–49 (2022)



Construction of Digital Twins Model for Logistics Finance

Guangren Chen^(✉)

School of Logistics, Guangdong Mechanic and Electrical Polytechnic, Guangdong, China
chgr263@sohu.com

Abstract. The current logistics finance model in China faces high financing costs and excessive operational risks, often leading to financial exclusion. Leaders in logistics finance need to integrate digital twins technology into their logistics finance operations. This involves systematically constructing a digital twins model for logistics finance, encompassing the design of the overall digital twins framework, logical models, and intelligent risk control mechanisms. This article introduces the systematic logic of digital twins in logistics finance, offering new insights and directions for the application of digital twins in the financial industry.

Keywords: Logistics Finance · Digital twins · Trusted Data · Smart Contract

1 Introduction

The logistics finance model highly integrates the financial activities of commercial banks with the operational characteristics of the logistics industry, integrating multiple functions such as financing and clearing, to provide customers with a one-stop solution [1]. However, in the actual business of logistics finance, banks, as lenders, need to carefully deal with legal flaws such as unclear ownership of inventory pledge rights, as well as liquidity and price risks caused by difficulties in realizing inventory and large fluctuations in market prices [2]. “Information opacity” is another key obstacle hindering the widespread application of logistics finance, which has significant negative effects, making banks face high costs and excessive risks in the process of searching for information. At this stage, the reasons for the insufficient application of logistics finance are multifaceted, including strict legal conditions, high credit costs, high degree of personalized services, high operating costs, limited selection of collateral, etc., but the root problem is information asymmetry. The final outcome of the game between lenders and borrowers is often the formation of logistics finance businesses with high financing costs and operational risks, and even frequent financial exclusion phenomena where banks refuse to lend due to cautious attitudes [3].

Under traditional financing models, banks find it difficult to verify the authenticity of specific information. This difficulty is manifested in their relatively limited understanding of the background of participating companies, inability to accurately assess the true value of collateral, and uncertainty about the authenticity of trade transactions.

This situation necessitates banks to extensively monitor the logistics, capital flow, and information flow at various stages throughout the entire supply chain. However, such monitoring incurs high costs, creates regulatory blind spots, and delays risk perception, leading to potential risks such as “false mortgages” and “multiple pledges on a single shipment” [4–6]. These issues can be addressed by integrating digital twins technology into logistics finance business processes, thereby innovating a new logistics financing model for the digital era. Essentially, this model combines various smart technologies, including sensors, the Internet of Things, warehouse accounting, and blockchain-based evidence storage [7]. However, the industry has yet to establish a standardized “digital twins of logistics finance” model.

Therefore, this paper proposes the construction of a “logistics finance digital twins” model with normative logic. The mechanism is that the financial platform uses digital twins technology to design the pledge financing scenario of logistics finance digital twins, provides real-time and accurate data matching for the whole process of pledge financing with warehouse receipts as the carrier, and realizes global optimization decision-making and risk control [8]. In this way, the operational logic of warehouse receipt pledge financing is reconstructed to achieve collaborative and intelligent decision-making, meet operational requirements such as insight into the situation, support decision-making, and closed-loop processing [9]. The logistics finance digital twins model uses the remote monitoring function of digital twins to map specific business scenarios in real time, penetrate to underlying assets, and provide remote intelligent monitoring and early warning services.

2 System Design of Logistics Finance Digital Twins Model

2.1 Conceptual Model of Digital Twins for Logistics Finance

The digital twins model architecture in the field of logistics finance is mainly constructed around four core elements: physical reality, virtual image, digital platform, and twins data [10]. It is driven by modern information technology innovation and cutting-edge service thinking to achieve comprehensive perception of the physical environment, facilities and equipment, human resources, and operational processes in the logistics system. Through this perception, the system accurately replicates a digital twins that highly matches the physical system in virtual space, enabling real-time data exchange and information synchronization between the physical system and the virtual twins [11].

The proposed Fig. 1 outlines a conceptual framework for a digital twins application in logistics finance, drawing upon the foundational five-dimensional models of digital twins. This framework encompasses three primary components: tangible elements existing within the physical realm; intangible counterparts residing in the digital domain; and the interlinking data and operations that bridge these dual realms.

Moreover, the logistics system’s actual and virtual data merge, creating a distinct dual dataset. This comprises extensive “twin big data” and granular “small data”. The twins big data serves for predictive analytics, performance assessment, and decision-making enhancement, whereas small data underpins the precise control and meticulous tweaking of logistical operations [12]. Such an integration model dramatically uplifts the intelligence quotient and operational efficacy of logistics finance.

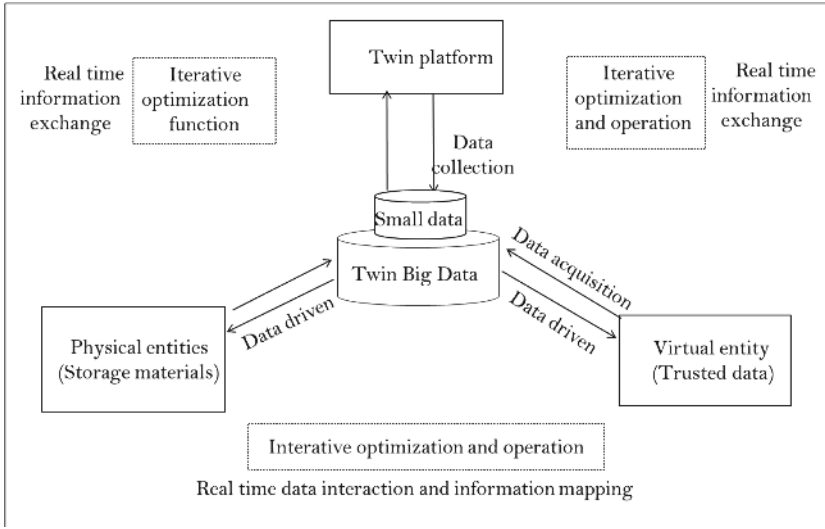


Fig. 1. Schematic Diagram of the Digital twins Concept Model.

With the help of data orientation, the digital twins platform can intelligently design and supervise the operation process of logistics finance, achieve real-time status monitoring and efficient management strategy optimization. It promotes the integration of all-round business, elements, processes and data between the physical world, virtual space and digital twins platform [13]. The purpose of establishing a digital twins system for logistics finance is to map a large amount of actual information to the virtual environment through efficient data collection of physical entities. Using virtual models, system developers can predict the behavior of logistics finance entities and apply these predictive information to entity decision-making. Due to the high matching and reliable data association between “entity-virtual”, it effectively solves the key problems in the field of logistics finance, such as information asymmetry and information authenticity verification [14].

2.2 Overall Framework of Digital Twins for Logistics Finance

Logistics finance is a comprehensive business field with multiple participants and activities, integrating various operational characteristics such as asset management, financial services, and transactions. The interactions among the participating entities are based on strict contractual relationships [3]. For banks, engaging in logistics finance business requires comprehensive processing of multidimensional information such as legal, accounting, transaction, and property rights identification. This information must be highly reliable, including its authenticity, integrity, immediacy, consistency, and security, to ensure the stable operation of the business [2]. Therefore, “trusted data” must be used to meet the data quality requirements that banks hope to achieve, and “data-driven”

to build a digital twins system structure. Due to the high similarity between “physical-virtual” and the connection of trusted data, the most critical data trust issue in logistics finance can be solved [4].

Figure 2 depicts the overall application framework of digital twins for logistics finance. This framework constructs a “physical-virtual” structure based on contract relationships, using trusted data, smart contracts, simulated virtual operations and data collection, consensus mechanisms, and parallel control of physical functional mechanisms. It builds an operational mechanism centered on bidirectional data interaction operations, ultimately achieving financial effects.

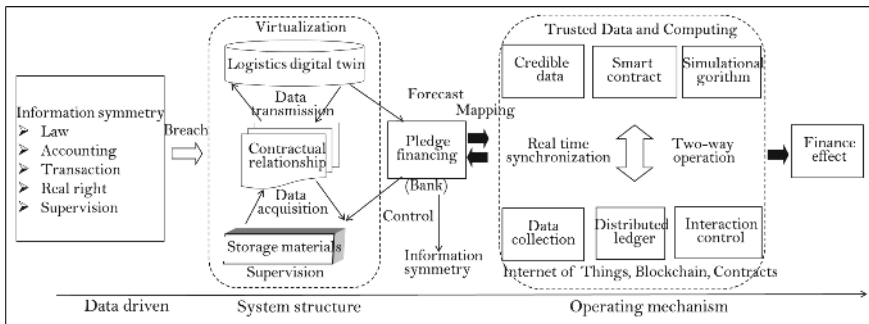


Fig. 2. Overall application framework of digital twins in logistics finance.

In the figure above, various contracts in logistics finance stipulate various rights, obligations, guarantee relationships, as well as multiple types of evidence, signatures, documents, and processes. Meeting the requirements of the contract is rigid and fundamental, and is a prerequisite for digital twins [2]. In the application scenario of adopting digital intelligent technology, contract relationships, as the core support of logistics finance, are also the basic rules for building digital twins.

The role and value of data twins is to achieve a true and comprehensive digital mapping based on the “contractual relationship” [15]. Unlike the “physical-virtual” two-in-one structure of other types of digital twins modeling, it is proposed to add a “contractual relationship entity” to construct a “three-in-one” structure combining “physical entity, contractual relationship entity, and virtual entity”, reflecting the specificity of logistics finance digital twins [16].

The “contractual relationship entity” is unique and very important. On the one hand, contract relations are different from geometry, physics, rules, etc. Most contract relations are implicit, non-visual, and even abstract logical relations—which can only be derived from logical reasoning based on factual relations [17]. For example, contract relations include important facts such as whether property rights/ownership have defects, whether transaction relations/pledge relations are established, and such contract relations are implied in a series of regulations, documents, and formal requirements such as signatures. The document processing function of traditional financing business is the key to clarifying the contractual relationship, which is the most fundamental, labor-intensive, and problematic link, contract relationship processing is the most difficult

business pain point for banks to break through, and there are technical difficulties in manually processing document issues which include:

1. A multitude of fraudulent documents exists, comprising counterfeit invoices, duplicated receipts, and manipulated signatures, necessitating effective countermeasures.
2. Document legitimacy encompasses the integrity of primary contracts, associated transport records, transactional warehouse and delivery papers, as well as inspection and acceptance certificates.
3. The harmony among diverse documents substantiates business credibility, by deterring conspiracies between upstream and downstream companies to fabricate invoices for non-existent transactions.
4. Documents undergo examination across processes, systems, departments, and institutions to ensure comprehensive validation.

In practice, it is difficult for banks to conduct in-depth investigations into the background of transactions. In most cases, they can only review the formal requirements, which naturally leads to deficiencies in the assessment of legal risk factors. It is seen that the main reasons for banks' difficulty in reviewing the true background of transactions are the low quality of information which is not timely and not comprehensive, insufficient evidential power, and coarse data granularity.

2.3 System Model of Logistics Digital Twins

Trustworthy data refers to high-quality data based on blockchain architecture, serving as the foundation for building digital twins. Data serves as the medium and fundamental element connecting physical entities with virtual entities. The fundamental requirements for data collection in logistics finance digital twins are to ensure real-time data and distributed arrangement [18]. Generally, the information that can be obtained by the digital twins model includes geometric dimensions, physical characteristics, and object behavior, and the data collection must meet the characteristics of real-time synchronization, reliable mapping, and high fidelity. However, logistics finance digital twins also superimpose consensus information and computational evidence on this foundation. Therefore, the functional structure of logistics finance digital twins is more complex.

High-level logistics financial data needs to be constructed with precise small data of physical entities, and such precise small data is consistent with the risk requirements of pledge financing. Financial technology integrates information technology, big data, etc. into the decision-making process of financial institutions to enhance their information screening, risk control, and other capabilities. Relying on the logistics, capital flow, information flow, etc. generated in the transaction process of the supply chain, which reflect the entire production and operation process of enterprises, such information is "small data", which is characterized by authenticity, accuracy, traceability, etc., meeting the data requirements of bank warehouse receipt financing business.

The participants in the logistics finance system have non-complete trust relationships, so the digital twins cannot adopt a centralized data access mode. It needs to be combined with blockchain to form a "decentralized" data governance mode to completely solve the information search problem. Blockchain obtains and stores evidence based on legal relationships and consensus mechanisms, thereby forming an evidence chain.

Based on the principles of evidence law such as original evidence, evidential force, and key evidence, a data fidelity mechanism and traceability mechanism are formed, thus constructing a digital mirror image that is highly similar to the physical entity.

Figure 3 depicts a versatile architecture where the logistics finance digital twins system is tailored to accommodate actual regulatory compliance and risk management needs.

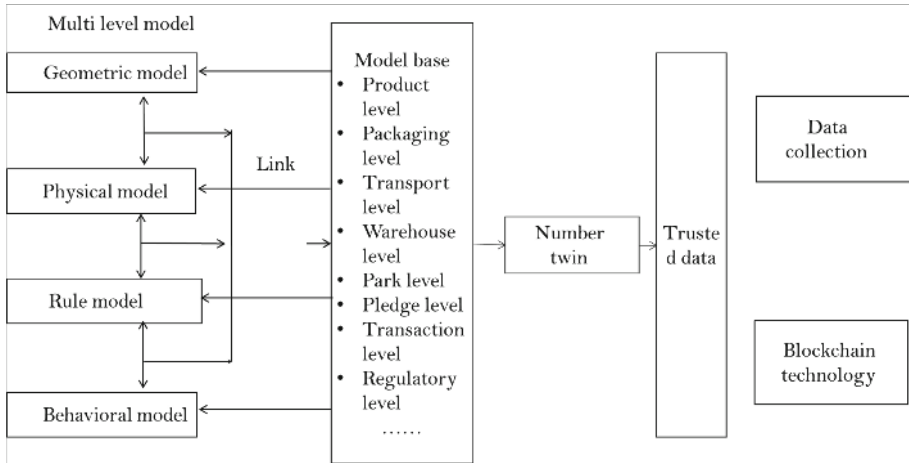


Fig. 3. Schematic diagram of logistics finance data twinning modeling.

This illustration showcases the distinctive operational features of the logistics digital twins, encompassing two-way data exchange processes and “tailored” oversight mechanisms.

Bidirectional Data Exchange Operation. Within the logistics finance framework, various entities necessitate distinct levels of real-time data exchange based on their unique requirements. Certain physical components exhibit static characteristics over extended periods, thereby necessitating no constant real-time synchronization during modeling and data acquisition, as long as their digital replicas accurately represent the present system architecture and specifications. Conversely, critical equipment, particularly those under surveillance, demand continuous and prompt monitoring data.

The trusted data generated through the logistics finance system is used to prove the facts related to logistics finance. On the one hand, the authenticity of the transaction is required, and on the other hand, the proof of facts and rights is necessary. Both are indispensable to generate digital trust. In digital twins, the high quality of information and strong collaboration among participants ensure that there is no loss or distortion of information transmission, which effectively guarantees the important value of using blockchain as the underlying architecture of digital twins.

Logistics finance digital twins are based on IoT smart devices, and may use methods such as data verification, big data, artificial intelligence, and edge computing to obtain

evidence, acquiring information used for property rights changes and physical movements. The authenticity of this information is guaranteed by encryption algorithms, and the chronological order of the information is determined by blockchain-pointing pointers, with timestamps recording the time the information was generated [19]. Facts are proven logically, using immutable, time-stamped data as evidence, ensuring that the data remains unchanged during its dissemination and storage processes, thereby generating digital trust.

The digitalization and clarification of contract relationships. Constructing a distributed application model with distributed identity authentication, distributed applications and databases (DAPP), forming a collaborative mechanism based on blockchain, and building a distributed organizational structure [20]. Using visualization technology to prove facts and transaction relationships from warehouse goods, documents to transaction data records and storage models, combining consensus mechanisms and visualization technology to clarify and explicit contract relationships, making digital contracts have functions such as property rights traceability, creditor's rights traceability, warehouse receipt circulation traceability, and processing flow. To achieve bidirectional interactive operation of data, the following principles need to be determined:

1. Modeling method. For the physical dimension, a geometric model is constructed based on the characteristics of assets such as shape, product packaging, transportation packaging, and storage location, using appropriate IoT data collection methods. For the behavioral dimension, a model of human/machine/equipment behavior characteristics is established based on data coupling and data articulation relationships related to handling, transportation, inbound and outbound operations, and warehousing.
2. Trustworthy data. The principle and method of collecting source data, combined with the duplex interaction mode (i.e., offline and online), decompose the scene into a hierarchical model (main chain, digital chain, auxiliary chain, physical chain), and analyze direct and indirect evidence collection methods and evidence effectiveness.
3. Data collection form, process, and method; bookkeeping method for trusted data formed by distributed ledger.

The “Specification” of Regulation. “Specification” is the most critical and unique functional mechanism of logistics finance. By analyzing numerous cases and examples that have occurred in the past, it can be found that asset pledge disputes in practice are mostly concentrated on the key issue of whether the pledge right takes effect, and the key determination basis is the “specification of the pledge”. Only when it is “specified” can the pledge right be established, which requires setting up a series of artificial arrangements for the movable property pledged to make it clearly distinguishable from other properties, and closely linked to the pledge contract, that is, it is necessary to study what kind of arrangements can achieve the purpose of factual proof.

The main specific methods include establishing clear physical identifiers, managing the pledged assets separately, setting up independent signs, and physically defining the pledged movable assets, such as warehouse partitions. The use of the Internet of Things can more effectively identify the specific characteristics of the pledged assets. Radio Frequency Identification (RFID) is physically combined with the pledged assets, which can effectively monitor and identify the pledged assets. In the digital twins, the specific

characteristics of the physical space must be reflected in data collection and consensus rules.

3 Risk Control of Digital Twins in Logistics Finance

The intelligent risk management of logistics finance digital twins essentially involves using the logistics finance digital twins as a parallel scenario and space for real-time interaction between the “virtual” and “physical” worlds, ultimately forming a closed-loop feedback.

“Bidirectional interaction and collaboration” are typical characteristics of digital twins, with the main goal of constructing a digital image that accurately maps the real world, transcending mere static historical records and enabling predictive foresight into future environmental evolution and the dynamics of the entity itself. Virtual twins achieve instant mapping between the physical entity and the digital world by monitoring sensor data of the physical entity in real time, subsequently conducting simulation operation tests in the virtual environment to further optimize control strategies and remotely manipulate the physical object. In the digital twins of the logistics and finance sector, bidirectional interaction encompasses various forms such as “physical-physical”, “virtual-virtual”, and “physical-virtual”, necessitating close collaboration among multiple elements within the logistics system, including personnel, machines, materials, and the environment. Particularly for “physical-physical” interaction, it facilitates effective communication, coordination, and joint task execution among physical devices, which is difficult to achieve in traditional risk control frameworks, effectively addressing many weaknesses in the “agent-principal” model. The “physical-virtual” interaction ensures real-time synchronization between the virtual model and logistics equipment, flexibly adjusting strategies according to virtual instructions, thereby enhancing response speed and efficiency.

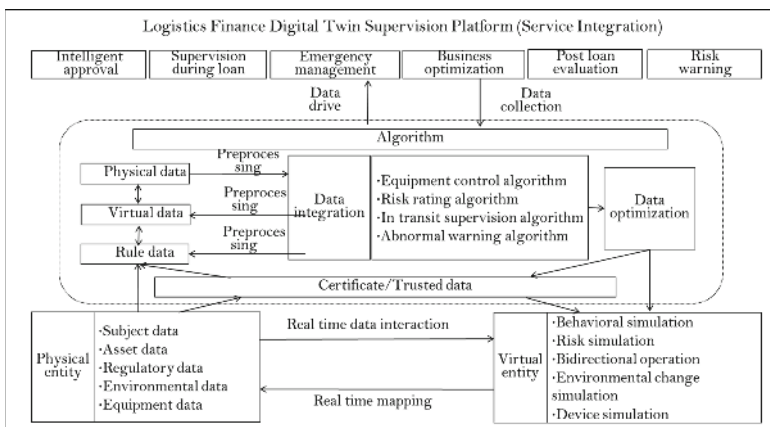


Fig. 4. Schematic diagram of intelligent risk management framework for logistics finance data twinsing.

Based on the above figure, logistics financial operators can establish an intelligent risk management mechanism for logistics financial data twins based on the principle of parallel risk control of smart contracts. Smart contracts are based on blockchain 2.0 technology, which uses algorithms and digital protocols to partially or fully perform centralized judicial functions, aiming to build a fair mechanism for the community of interests that guarantees procedural fairness, outcome fairness, and factual fairness [21]. This innovative mechanism is essentially a reshaping of judicial functions, firmly rooted in the principles of the rule of law, emphasizing the accurate identification of facts and the decisive role of evidence in determining facts [22]. Through technological means, smart contracts can automatically perform a series of links such as evidence collection, fact judgment, and legal ruling, thus ensuring the fair adjudication and distribution of rights and interests. Its built-in compulsory execution and automatic execution characteristics effectively curb unfair behaviors that may arise in the execution of traditional contracts due to personal self-interest. System function interaction and integration:

1. **Technological Infusion:** Streamlined regulation compliance in banking risk management is achieved through the implementation of Internet of Things technology and advanced algorithms, effectively substituting human decision-making processes.
2. **Model Synthesis:** The development, assessment, and validation of virtual entity mapping models involve correlation, mapping, and their seamless integration.
3. **Information Integration:** Real-time, reliable, and predictive data combine to vividly illustrate the evolving dynamics and patterns of system performance, component behavior, and more.
4. **Seamless Service Coordination:** Driven by data and supported by diverse algorithms, the synergy between data fusion and risk control replaces manual operations, offering precise control and decision-making assistance for both physical and virtual system dynamics.

Based on the integration framework of “physical integration, model integration, data integration, and service integration”, Fig. 4 proposes an intelligent risk management system framework.

4 Conclusion

The evolution of collateralizing movable assets has transitioned from initial notions of logistics banking to contemporary conceptions like digital logistics finance. This approach fundamentally leverages information technology to mitigate information disparities, minimize operational expenses, and escalate the sophistication of risk management intelligence.

Digital twins symbolizes a novel phase in the journey of industrial digitization. Establishing a logistics finance digital twins grounded in reliable data serves as the bedrock for mapping out legal associations and facilitating data exchange within digital twins. Unlike its application in design and manufacturing, the logistics finance digital twins embodies financial and transactional characteristics, intrinsically tied to contractual agreements and asset ownership.

The “physical-virtual” connection is established through trusted data, and the purpose of constructing a logistics finance digital twins system is to fully utilize the data

acquisition function of physical entities, mirror a large amount of real information to virtual entities, use virtual entity models to predict the behavior of physical entities, and feed the predicted information back to physical entities. Contractual relationships serve as the core -essence and core rules for constructing digital twins. For the intelligent risk management of logistics finance digital twins, it essentially involves using logistics finance digital twins as parallel scenarios and their parallel spaces, with bidirectional real-time interaction between “virtual” and “physical” entities, forming a closed-loop feedback.

References

1. Zhang, Y., Li, X., Wang, L., Zhao, X., Gao, J.: Financing capital-constrained third party logistic firms: fourth party logistic driven financing mode vs. private lending driven financing mode. *Int. J. Prod. Res.* **60**(9), 2963–2982 (2021)
2. Xu, X.: Warehousing financing behavior of global logistics system under the situation of financial crisis. In: Yang, Y.; Mi, C.; Zhao, L., Lam, S. (eds.) *Global Topics and New Trends in Coastal Research: Port. Coastal and Ocean Engineering*. Journal of Coastal Research, Special Issue No. 103, pp. 658–662 (2020)
3. Li, M., Shao, S., Ye, Q., Xu, G., Huang, G.Q.: Blockchain-enabled logistics finance execution platform for capital-constrained E-commerce retail. *Robot. Comp.-Integr. Manuf.* **65**, 101962 (2020)
4. Lu, B., Lu, H., Wang, H.: Design and value analysis of the blockchain-based port logistics financial platform. *Maritime Policy & Management*, 1–25 (2023)
5. Wang, A., He, Z., Wang, Y.: RETRACTED: Risk Assessment of Logistics Finance Enterprises Based on BP Neural Network and Fuzzy Mathematical Model **1**, 5915–5925 (2020)
6. Li, T., Fang, W., Desheng, D. W., Zhang, B.: Inventory financing a risk-averse newsvendor with strategic default. [Inventory financing a risk-averse newsvendor] *Industrial Management & Data Systems* **120**(5), 1003–1038 (2020)
7. Zhu, Y., et al.: Production logistics digital twins: Research profiling, application, challenges and opportunities. *Robot. Comp.-Integr. Manuf.* **84**, 102592 (2023)
8. Pan, Y., Zhong, R.Y., Qu, T., Ding, L., Zhang, J.: Multi-level digital twins-driven kitting-synchronized optimization for production logistics system. *Int. J. Prod. Econ.* **271**, 109176 (2024)
9. Pan, Y.H., Qu, T., Wu, N.Q., Khalgui, M., Huang, G.Q.: Digital twins based real-time production logistics synchronization system in a multi-level computing architecture. *J. Manuf. Sys.* **58**(B), 246–260 (2021)
10. Guo, D., Zhong, R.Y., Rong, Y., Huang, G.G.Q.: Synchronization of Shop-Floor Logistics and Manufacturing Under IIoT and Digital twins-Enabled Graduation Intelligent Manufacturing System. *IEEE Transactions on Cybernetics* **53**(3), 2005–2016 (2023)
11. Jeschke, S., Grassmann, R.: Development of a generic implementation strategy of digital twins in logistics systems under consideration of the German rail transport. *Appl. Sci.* **11**, 10289 (2021)
12. Rigó, L., Fabianová, J., Lokšík, M., Mikušová, N.: Utilising digital twins to bolster the sustainability of logistics processes in industry 4.0. *Sustainability* **16**, 2575 (2024)
13. Nicola, B., Serena, F.: Digital twins and human factors in manufacturing and logistics systems: state of the art and future research directions. *IFAC-Papers OnLine* **55**(10), 1893–1898 (2022)
14. Uhlenkamp, J.F., Hauge, J.B., Broda, E., Lütjen, M., Freitag, M., Thoben, K.D.: Digital twins: a maturity model for their classification and evaluation. *IEEE Access* **10**, 69605–69635 (2022)

15. Neugebauer, J., Heilig, L., Voß, S.: Digital twins in the Context of Seaports and Terminal Facilities. *Flex Serv. Manuf. J.* (2024)
16. Jernej, P., Miha, P., Hugo, Z., Matevz, R., Marko, S., Niko, H.: Edge computing and digital twins based smart manufacturing. *IFAC-Papers OnLine* **54**(1), 831–836 (2021)
17. Aretoulaki, E., Ponis, S.T., Plakas, G., Tzanetou, D.: (DT4Smart) discrete event simulation and Digital twins in warehouse logistics: a bibliometric and content analysis-based systematic literature review. *Int. J. Comp. Integr. Manuf.*, 1–28 (2024)
18. Li, Y., Chen, M., Li, Q., Zhang, W.: Enabling multilevel trust in privacy preserving data mining. *IEEE Trans. Knowl. Data Eng.* **24**(9), 1598–1612 (2012)
19. Habib, S.J., Marimuthu, P.N.: Development of analytical model for data trustworthiness in sensor networks. 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), pp. 467–472. Crans-Montana, Switzerland (2016)
20. Park, J., Jeong, S., Yeom, K.: Smart contract broker: improving smart contract reusability in a blockchain environment. *Sensors* **23**, 6149 (2023)
21. Hu, K., Zhu, J., Ding, Y., Bai, X., Huang, J.: Smart contract engineering. *Electronics* **9**, 2042 (2020)
22. Gunnar, P.: Smart contracts for smart supply chains. *IFAC-Papers OnLine* **52**(13), 2501–2506 (2019)

Data Mining and Machine Learning



The Study of Financial Risk Analysis and Prediction for Enterprises Based on Big Data Mining

Min Jiang^(✉)

Institute of Accounting, Chongqing Business Vocational College, Shapingba,
Chongqing 401331, China
jmmjmmjmm315@163.com

Abstract. Based on big data mining technology, this paper studies the construction and application of enterprise financial risk analysis and prediction model. By comparing multiple machine learning algorithms, this paper constructs an efficient financial risk prediction model and applies it to the risk analysis of enterprise financial data. The research results show that support vector machine (SVM) performs best in evaluation indicators such as accuracy and F1 score, while random forest and neural network can also provide good prediction results in certain specific cases. Through visualization methods such as box plots, pairwise relationship diagrams and violin plots, the intrinsic relationship between financial indicators and enterprise financial risks is further revealed. This study provides scientific decision-making support for enterprise financial management and demonstrates the wide application potential of big data technology in the field of risk prediction.

Keywords: Financial Risk Prediction · Machine Learning · Support Vector Machine · Risk Analysis · Big Data Mining

1 Introduction

1.1 Research Background and Importance

Meanwhile, during the acceleration period of world economic integration, the financial risks in front of enterprises have obviously tended to be more diversified with their complex characteristics. How to detect and predict these risks is among the core issues of corporate finance management. Financial risk is related to the profitability and survival and development of enterprises, as well as to a big influence on the market competitiveness and the realization of long-term strategic goals. Traditional financial risk analysis methods are basically dependent on manual experiences and some basic statistical models; these methods are generally not able to capture in real time and comprehensively some early warning signals of risk hiding behind complex financial data [1]. Big data technology is developing rapidly, and it can collect and process a large amount of financial data for enterprise use, thus providing new opportunities for financial risk prediction and management. In financial risk analysis, big data mining technology, especially the

machine learning algorithm, is able to find out potential patterns and associations through deep analysis of large-scale historical data, thus providing more accurate support for risk warning. For example, some machine learning methods like support vector machines and random forests have been successfully applied in many areas with their excellent predictive ability, and their applications are also increasing in the financial field [2].

Due to high dimensionality and nonlinear, the nature of financial data affects the classic methods of analysis. Big data mining technology can introduce a deep understanding of financial data and extract more complex nonlinear relationships, which improves the accuracy and robustness of the prediction model. Machine learning technology is especially good at processing such complex data and can make relatively accurate future risk predictions based on the changing trends of historical data. Therefore, by applying big data in establishing a corporate financial risk prediction model, it would not only optimize the enterprise financial decision-making process but also reduce financial risks and lower uncertainty in losses effectively, which has very important value in theory and practice [3].

1.2 Research Objectives

The key purpose of this paper is to explore how to leverage big data mining technology for better analysis and prediction of corporate financial risks. First, an effective model for the financial risk prediction is constructed through a comparison of several common machine learning algorithms. In a word, at its very core is the accurate identification of the risk factors from learning in historical financial data, after which comes a scientific solution on the enterprise risk assessment and early warnings. In this paper, the authors have selected algorithms like SVM, random forest, and neural network for deep digging into the financial data of corporations in order to compare the performance of each model for different evaluation indicators. Second, it introduces some data visualization technologies like box plots, pairwise relationship diagrams, and violin plots that provide an in-depth reveal of the intrinsic relationship between financial indicators and corporate financial risks. This not only intuitively displays the distribution of various financial data, but also intuitively provides a basis for financial managers to make decisions. In the end, through the research, it can provide enterprises with a comprehensive, accurate, and easy-to-operate financial risk prediction tool, and demonstrate the wide application potential of big data technology in the field of financial risk prediction.

2 Literature Review

2.1 Relevant Theories of Corporate Financial Risk

The theoretical framework of corporate financial risk management includes a series of theoretical models. At the core, these theories' function is to identify, evaluate, and control risk. The most well-known classic theories of financial risks include Modern Portfolio Theory (MPT) and Capital Asset Pricing Model (CAPM), which have been applied in corporations for a long period as guidance for making financial decisions. The former mainly tells how diversified investment could lower risks, while the latter

focuses on market risks and returns on corporate assets [4]. However, the economic environment is complicated, corporate financial management becomes more and more sophisticated; traditional theoretical methods are incompetent while facing large-scale data and complex risk factors [4]. In recent years, behavioral economics and complex system theory have been gradually brought into financial risk management research. These theories emphasize the diversity and uncertainty of risks and are of the view that their emergence is often entwined with multiple factors inside the system, so what is needed is a more comprehensive and dynamic risk prediction model. These theories provide an important theoretical basis for financial risk assessment and management and promote more risk prediction methods based on data analysis [5].

2.2 Application of Big Data Mining in Financial Risk Analysis

In recent years, the technologies of big data mining have attained a large number of developments in corporate financial risk analysis. Big data technology would therefore be a useful instrument for financial risk predication by gathering massive data, processing it, and analyzing it. Following this, machine learning models, among them support vector machines or decision trees and random forests, already realized the successful application around prediction studies on the financial risks [6]. These methods can learn the potential pattern of historical data and provide an efficient early warning mechanism when facing unknown risks. Meanwhile, some data visualization technologies, such as heat maps, box plots, and violin plots, are also widely used to display various dimensions of financial risks and help financial managers find and analyze potential risks more intuitively. Furthermore, the big data-based risk analysis model should be updated and adjusted in real-time so that the accuracy and timeliness of the forecasting results can be further improved [7]. Besides, with further development, the artificial intelligence technology will develop a deep learning model in a way that it applies strong application potentials in the analysis of financial risks, so that financial forecasting and a risk management strategy can be further optimized.

2.3 Deficiencies and Challenges of Current Research

Although big data mining technology has been widely used in financial risk analysis, it still suffers from a series of key challenges and deficiencies. First, the high dimensionality and nonlinear relationships in financial data add more difficulty to modeling and analysis. Although some machine learning algorithms can already process high-dimensional data, how to effectively avoid the curse of dimensionality and improve the computational efficiency of the algorithm remains an urgent problem to be solved [8]. Second, data quality and data integration issues are also important factors that restrict the wide application of big data technology. Financial data is generated from so many sources and in such a variety of forms that cleaning and integrating the data is very time-consuming and complicated work [9]. Besides, although deep learning has shown its powerful capability in financial risk prediction, the black box nature makes the model less interpretable, which may lead to a situation where the decision makers do not have sufficient trust in the prediction results in practical applications [10]. Therefore, how

to improve the interpretability and transparency of the model remains an important direction of current research.

3 Financial Risk Analysis Methods and Models

3.1 Basic Formulas and Models for Financial Risk Analysis

This diagram (Fig. 1) represents the main informational flow of a financial risk prediction model, emphasizing the iterative process of optimization at each stage in model creation. It starts with pre-processing, then moves on to feature selection, training, and risk scoring in a process refined through continuous feedback.

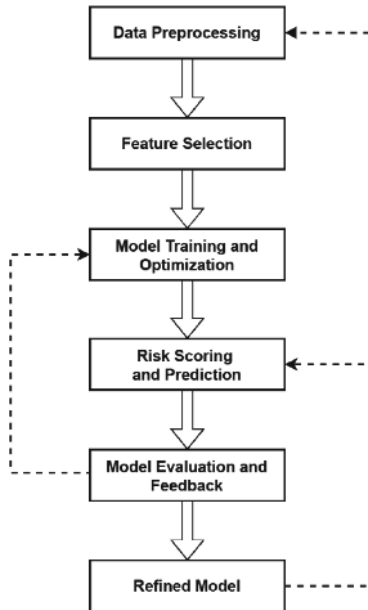


Fig. 1. Financial Risk Prediction Model Information Flow

Financial leverage ratio measures the relationship between corporate liabilities and shareholders' equity.

$$D/E = \frac{\text{Total Debt}}{\text{Total Equity}} \quad (1)$$

Among them, D/E is the leverage ratio, which indicates the proportion of debt used by the enterprise relative to shareholders' equity.

Current ratio: reflects the short-term debt repayment ability of the enterprise.

$$\text{Current Ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}} \quad (2)$$

Z-score model: predicting the possibility of corporate bankruptcy.

$$Z = 1.2 \cdot X_1 + 1.4 \cdot X_2 + 3.3 \cdot X_3 + 0.6 \cdot X_4 + 1.0 \cdot X_5 \quad (3)$$

Among them, X_1, X_2, X_3, X_4, X_5 are the weighted values of different financial indicators. The lower the Z value, the higher the bankruptcy risk.

3.2 Design and Optimization of Financial Risk Prediction Model

Support vector machine (SVM): classifies financial risks by maximizing the margin.

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (4)$$

Among them, $\mathbf{w}$ and b are the parameters of the hyperplane, and SVM distinguishes financial risk categories by maximizing the interval.

Neural network model simulates the connection between neurons to predict risks.

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (5)$$

Among them, w_i is the weight of the input feature, x_i is the input feature, b is the bias term, and f is the activation function, usually a nonlinear function is used.

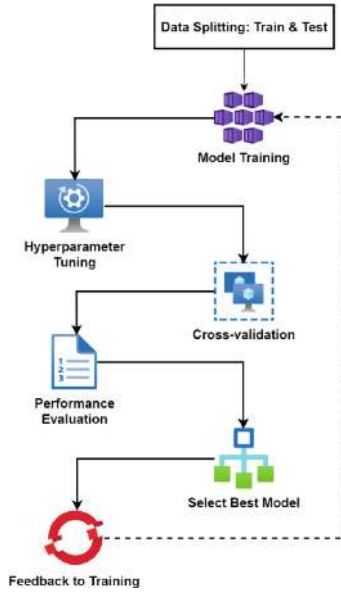


Fig. 2. Model Training and Evaluation Process

Figure 2 shows the closed-loop process of model training and evaluation. First, training and testing are performed through data partitioning, then cross-validation and hyperparameter tuning are used to optimize the model, and finally performance evaluation is performed to select the best model. If the evaluation result is not ideal, the model will re-enter the training phase for optimization based on the feedback information, thus forming a continuous improvement cycle.

4 Evaluation and Application of Financial Risk Prediction Models

4.1 Model Evaluation Indicators and Methods

Some common evaluation indicators adopted in the process of financial risk prediction model evaluation are accuracy, recall, F1 score, ROC curve, and AUC value. The following box plot will help one to be able to show deeper knowledge about model performance in different manners of financial risk prediction. We based it on an open financial statement data dataset drawn from 500 companies distributed in various industries, considering 3,000 statements of the same. Information gathered will include key financial data covering company financial ratios, which define revenues, liabilities, and profits.

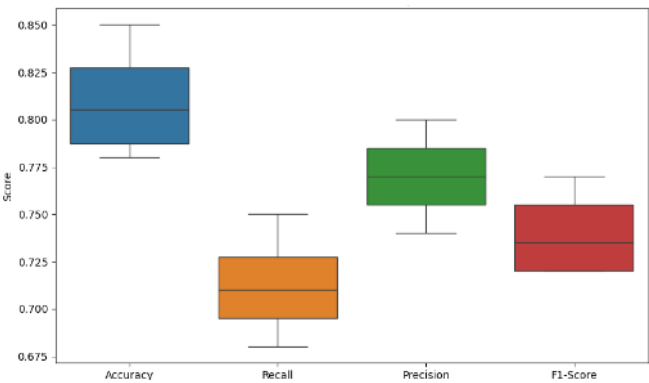


Fig. 3. Model Performance Comparison

The distribution of the main performance metrics (Fig. 3), Accuracy, Recall, Precision, and F1-Score, of the machine learning models SVM, Logistic Regression, Random Forest, and Neural Network. One can observe from this graph that while SVM generally yields higher scores across most metrics, other models, like Random Forest, are doing pretty well in certain areas.

4.2 Analysis and Interpretation of Risk Prediction Results

In actual risk prediction, we use different indicators to evaluate the effectiveness of financial risk models. The following is an analysis of the relationship between financial data indicators and predicted risks using a pairwise relationship diagram.

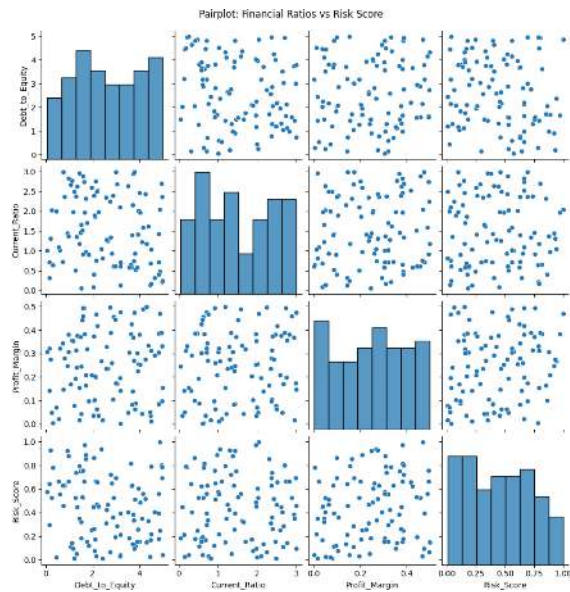


Fig. 4. Financial Ratios vs Risk Score

The pairplot (Fig. 4) above shows the scatter relationships between various financial ratios (Debt to Equity, Current Ratio, Profit Margin) and the predicted Risk Score. It allows for a quick visual inspection of how these financial metrics correlate with the predicted risk, helping to identify patterns or anomalies in the data.

4.3 Practical Application and Effect of the Model

In actual use of the financial risk prediction model, except for the standard evaluation indices-based model effect analysis, the violin plot should be used to show the different distribution of different companies about their financial risks. The following is the violin plot used to show the distribution of different companies' financial risk.

This violin plot (Fig. 5) depicts the distribution of scores on financial risk for five different companies. Each "violin" shows the range of risk scores a company has, with wider parts showing higher density of risk scores. The following visualization provides insights into the variability of financial risk across companies, in order to understand which companies may exhibit greater financial instability or volatility.

4.4 Results Summary

As shown in Table 1, there were certain differences in the predicting effect among different models due to the analysis of evaluations and applications. According to all the accuracy, recall, precision, F1 score, and other results using various indicators, the SVM model performed well in most situations, considering the better prediction ability given, while the random forests and neural networks are performed in some indicators.

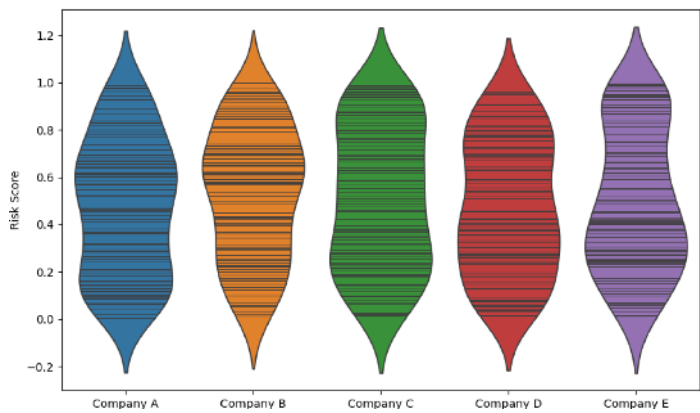


Fig. 5. Financial Risk Distribution Across Companies

Besides, by comparing the pairwise relationship diagrams and violin plots, the following can be further revealed: the close relationship between different financial indicators and risk prediction results, the differences in financial risks among companies. The above analysis provides strong support for accurately predicting financial risks and performing enterprise risk management.

Table 1. Model Performance Metrics and Financial Risk Scores

Metric	SVM	Logistic Regression	Random Forest	Neural Network
Accuracy	0.85	0.78	0.82	0.79
Recall	0.75	0.70	0.72	0.68
Precision	0.80	0.74	0.78	0.76
F1-Score	0.77	0.72	0.75	0.72
Company A	0.85	0.78	0.90	0.76
Company B	0.65	0.70	0.60	0.72
Company C	0.50	0.55	0.58	0.62
Company D	0.92	0.89	0.87	0.93
Company E	0.80	0.85	0.78	0.83

5 Conclusion and Future Work

5.1 Main Conclusions

This paper will conduct an intensive discussion on the establishment and application of financial risk prediction models, using big data mining methods to analyze the financial risk status of enterprises. To begin with, it is found in this study that financial indicators

bear an important role in predicting corporate financial risks. Through the comparison of several machine learning algorithms, it can be seen that the support vector machine algorithm has better performance in both accuracy and F1 score and is thus the most promising model for financial risk prediction. Besides, under specific conditions, random forest and neural network models also can provide competitive prediction results, especially when the scale of data is large, the performance is relatively stable. Such a complicated relationship between financial ratios and forecasted risks can be visualized by techniques like box plots, violin plots, and pairwise relationship diagrams; therefore, such visualization gives the important basis of decisions on both financial management and risk monitoring.

Second, in the paper, financial data from 500 companies are analyzed to prove that the model is effective in practice. It will be very clear to compare the differences in the financial health status of different companies with different financial risk scores, thus providing scientific support for enterprise financial decisions. The successful application of the model indicates great potential for big data technology in financial risk prediction, especially in processing massive corporate financial data and unstructured data. In general, this study provides effective tools and methods for corporate risk management and capital market analysis.

5.2 Research Limitations

Although the presented study has achieved a series of results in constructing and applying the financial risk prediction model, it also has some limitations: firstly, although the dataset applied in this study covers a large number of industries, relatively small sample sizes limit diversity, which may further affect the applicability of the model in specific industries or enterprise types. The influences from such aspects as industry characteristics and regional economic environment mean that financial data could only ensure the full applicability of financial risk prediction in special industries or small enterprises. While we have tried various kinds of machine learning algorithms, we did not take all the possible models that might be applied, for example, combining some deep learning and ensemble learning methods. There may be more effective algorithms that further improve the accuracy of the prediction of the model.

5.3 Future Research Directions

For future research, several directions might be considered. First, further expansion of the scale and diversity of the data set will cover more industries and regions, enhancing the generalization ability of the model. It is also possible to study how different kinds of data, such as the integration of financial statements, market data, socio-economic data, and unstructured data like news, social media content, etc., can be put together for a comprehensive improvement in the accuracy of financial risk prediction with the use of multimodal data mining. The result will ensure high-precision risk prediction with multiple sets of decision supports during corporate management.

On the other side, more powerful deep learning methods and reinforcement learning techniques would be another key perspective worthy of future research. In particular, some deep learning techniques, like DNN and CNN, would perform well in dealing with

high-dimensional complex data with good performance. The reinforcement learning method will combine further for improving adaptability of model in a complex dynamic environment with the goal to guide companies to conduct real-time risk assessment when using real-time decision-making processes. Besides, another important direction for future research is to investigate some integrated learning methods such as XGBoost, LightGBM, etc. Those methods can optimize the performance of prediction and reduce overfitting risks through multimodel fusion.

Acknowledgement. The key funding project of the 14th Five-Year Plan for Education and Science of Chongqing in 2021 (No. 2021-GX-057) “Research on the connotation and practice Path of Apprenticeship with Chinese Characteristics under the guidance of Model Worker Spirit, labor spirit and craftsman Spirit in the New Era” phased research results.

References

1. Jiang, L., et al.: Diabetes risk prediction model based on community follow-up data using machine learning. *Prev. Med. Rep.* **35**, 102358 (2023)
2. Liu, C., et al.: Distant metastasis pattern and prognostic prediction model of colorectal cancer patients based on big data mining. *Front. Oncol.* **12**, 878805 (2022)
3. Parikh, R.B., Gdowski, A., Patt, D.A., Hertler, A., Mermel, C., Bekelman, J.E.: Using big data and predictive analytics to determine patient risk in oncology. *Am. Soc. Clin. Oncol. Educ. Book* **39**, e53–e58 (2019)
4. Xie, X., et al.: Evaluating cancer-related biomarkers based on pathological images: a systematic review. *Front. Oncol.* **11**, 763527 (2021)
5. Wang, H., et al.: A cost-effective machine learning-based method for preeclampsia risk assessment and driver genes discovery. *Cell Biosci.* **13**, 41 (2023)
6. Shi, S., et al.: Predictable regulation of survival by intratumoral microbe-immune crosstalk in patients with lung adenocarcinoma. *Microb. Cell* **11**, 29–40 (2024)
7. Eom, G., Byeon, H.: Development of keyword trend prediction models for obesity before and after the COVID-19 pandemic using RNN and LSTM: Analyzing the news big data of South Korea. *Front. Public Health* **10**, 894266 (2022)
8. Zheng, P., et al.: Establishment of a prognostic model of lung adenocarcinoma based on tumor heterogeneity. *Front. Mol. Biosci.* **9**, 807497 (2022)
9. Shimpi, N., Glurich, I., Panny, A., Hegde, H., Scannapieco, F.A., Acharya, A.: Identifying oral disease variables associated with pneumonia emergence by application of machine learning to integrated medical and dental big data to inform eHealth approaches. *Front Dent Med* **3**, 1005140 (2022)
10. Wang, Y., et al.: Assessment of the impact of geogenic and climatic factors on global risk of urinary stone disease. *Sci. Total. Environ.* **721**, 137769 (2020)



Mean-CVaR Portfolio Model Based on Graph Neural Network

Guogang Zheng^(✉)

School of Information, Huazhong Agricultural University, Wuhan, China

ggz_1@webmail.hzau.edu.cn

Abstract. With huge risks in financial investment, how to efficiently transform from subjective investment to scientific quantitative investment is particularly crucial. For this reason, this paper proposes a mean-CVaR portfolio model based on graph neural network (GNN), aiming at improving the scientificity and efficiency of investment decisions. The variance of future stock returns is predicted by graph convolution network and compared with XGBoost and LSTM models. Then, this research further uses CVaR to quantify investment risk and builds a financial portfolio model based on mean-CVaR. The stock data in the field of new energy vehicles are selected for prediction and analysis to solve the optimal investment portfolio. The research results confirm the feasibility and superiority of the model in portfolio management.

Keywords: Graph Neural Network · Portfolio Optimization · Financial Investment · Mean-CVaR

1 Introduction

With the increasing globalization of economic activities, portfolio management in contemporary financial markets is not only an investment technique, but also a vital means for investors to seek stable returns and control risks in complex market environments [1]. Since Harry Markowitz proposed the mean-variance model in 1952 [2], which has laid the foundation for portfolio theory. However, it has limitations, such as relying on a large number of parameters, assuming that returns follow a normal distribution, and using only variance to measure risk [3]. Based on Markowitz's work, later scholars loosened the hypothesis of the original model for optimization [4–6].

Compared with traditional statistical methods, machine learning methods can more accurately predict key parameters, build more robust investment strategies and realize algorithm optimization [7]. Some scholars apply machine learning models to predict market risk. For example, Chen et al. (2016) proposed the XGBoost algorithm, improved the shortcomings of the GBDT algorithm, and verified that XGBoost outperforms SVR in stock forecasting [8]. Wang Y and Guo Y (2019) optimized the XGBoost model with the grid search algorithm to construct the GS-XGBoost financial forecasting model, which is applied to short-term stock forecasting [9]. Sun Y (2020) constructed a stock selection strategy centered on a random forest and support vector machine (SVM), which indicates

that the stock selection strategy based on SVM is more accurate than the random forest method [10].

With the widespread application of deep learning, academia and industry have begun to explore the use of deep learning models to predict the expected rate of return and variance of financial products, so as to optimize asset allocation and risk control strategies [11]. In finance, deep learning applications include models such as feedforward neural networks (FNN), convolutional neural networks (CNN), and long-short-term memory networks (LSTM) based on recurrent neural networks (RNN) [7, 12, 13]. However, the input of these models is usually only a feature matrix, failing to fully consider the allocation of existing portfolios, which may affect the accuracy of the model's evaluation of transaction costs caused by portfolio adjustments [14]. Graph neural network (GNN) provides a more effective solution for portfolio management because of its advantages in analyzing complex relationships between financial assets [15].

With its powerful learning capabilities and high data-driven performance, deep learning has made significant progress in many fields. GNN has been a popular research topic in recent years [16, 17]. With the development of Internet technology, non-Euclidean data forms (such as communication networks, social networks, and stock-related networks in financial markets) have become more and more common. Traditional CNNs or RNNs cannot effectively process these graph data [18]. Graph neural networks are specially designed for analyzing and processing such complex graphic data, which shows unique advantages in processing non-Euclidean spatial data. In this paper, the graph convolution network (GCN) is used to predict the future variance of return. On this basis, the optimal portfolio is constructed to minimize the variance of return.

To sum up, based on GNN and CVaR risk value, this paper proposes a GNN-based mean-CVaR portfolio model. The variance of stock return is predicted by GCN and compared with XGBoost and LSTM model. Then, CVaR is used to quantify investment risk and construct a portfolio model based on mean-CVaR. According to the results, the optimal portfolio constructed by the GCN model is better than other models in terms of risk resistance and economic benefits.

2 Portfolio Model Based on CVaR Risk

CVaR refers to the average loss when the investment loss exceeds the extreme value of VaR for a certain period in the future under a certain level of confidence. It is proposed to make up for the deficiency of VaR, with its formula as follows:

$$\text{CVaR}_\alpha(\gamma) = E(\gamma | \gamma \geq \text{VaR}_\alpha(\gamma)) = \frac{1}{1 - \alpha} \int_{f(x,y) \geq \mu_\alpha(x)} f(x,y) p(y) dy \quad (1)$$

where γ is the portfolio loss, $f(x,y)$ is the profit and loss function of the investment portfolio, X is the feasible set of the investment portfolio, x is the weight vector of each asset in the investment portfolio, and y is the market price vector.

The distribution density function for any given x and y is $P(y)$, assuming y is a continuous random variable. For any $\mu \in R$, there is:

$$\psi(x, \mu) = \int_{f(x,y) \leq \mu} p(y) dy. \psi(x, \mu) \quad (2)$$

$\psi(x, \mu)$ is a cumulative probability distribution function with incrementing μ . For any $\alpha(0, 1)$, it is defined as:

$$\begin{aligned} \mu_\alpha(x) &= \min\{\mu \in R : \psi(x, \mu) \geq \alpha\}, \\ \varphi_\alpha(x) &= E(f(x, y) | f(x, y) \geq \mu_\alpha(x)) = \frac{1}{1-\alpha} \int_{f(x,y) \geq \mu_\alpha(x)} f(x, y) p(y) dy, \end{aligned} \quad (3)$$

Then $\mu_\alpha(x)$ and $\varphi_\alpha(x)$ are VaR and CVaR with a confidence level of α .

Based on the mean-variance model and CVaR method, the mean-CVaR model is established as follows:

$$\min_{x, \theta} \hat{F}_\alpha(x, \theta) = \theta + \frac{1}{J(1-\alpha)} \sum_{k=1}^J d_k \quad (4)$$

$$\text{s.t.} \begin{cases} x^T K \leq k_0 \\ d_k \geq -x_k r_k - \theta \\ d_k \geq 0 \\ \sum_{i=1}^n x_i = 1 \\ 0 \leq x_i \leq 1, \end{cases} \quad (5)$$

where the rate of return r_k and forecast variance K are the input parameters. Confidence α , the number of assets n , the number of return scenarios J , and the level of variance k_0 are given parameters. Asset weights x , dummy variables d_x and VaR values θ are the model outputs.

3 Stock Forecast Based on GCN

GNN is a deep learning model designed for processing graph data structures. In graph data, nodes represent entities, and edges describe relations between entities. GNN combines node features with an adjacency matrix and iteratively updates node states to achieve message delivery and feature aggregation on graphs. Through embedding propagation, GNN aggregates the neighbor information of the target node to obtain high-order neighborhood information and updates the representation of the target node to finally obtain the node representation [19].

GCN is a variant of GNN, which captures the complex interaction of graphs by convolving nodes and edges of graphs. In 2013, Bruna et al. proposed to divide the GCN into spatial graph convolution and spectral domain graph convolution [20]. In 2017, Thomas Kipf et al. [21] improved the convolution operation based on ChebNet and introduced a series of operations such as re-normalization. Then, they extended the convolution to the multi-dimensional space of the eigenmatrix, which finally obtained a commonly used GCN model formula:

$$y = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta\right) \quad (6)$$

3.1 Stock Data Collation

This research selects nine stock data in the field of new energy vehicles for correlation analysis. To construct the GCN model, this paper collects and preprocesses the data in detail, aiming to predict the future return variance of these stocks through the GCN model. On this basis, an optimal portfolio model is built.

Data Collection

This paper collects the daily stock data of nine new energy vehicle companies, including Tesla (TSLA) and Weilai Automobile (NIO) from public financial databases. The periods of data range from January 3, 2022, to December 29, 2023. In addition, these companies are important representatives in the new energy vehicle industry with high market liquidity and scale. The dataset includes information such as opening price, highest price, lowest price, closing price, trading volume and turnover, with the specific company information shown in Table 1.

Table 1. Constituent Stocks of the New Energy Vehicle Industry and Their Stock Codes

Company	Stock Code	Company	Stock Code
TSLA	TSLA.O	NKLA	NKLA.O
FSR	FSR.N	RIVN	RIVN.O
LCID	LCID.O	VFS	VFS.O
LI	LI.O	XPEV	XPEV.N
NIO	NIO.N		

Data Preprocessing

To meet the input requirements of the GCN model, this paper conducts the following preprocessing operations on the data.

Step1: Ensure the transaction time is listed as a date type and set as an index, then use the time interpolation method to fill in the missing values and reset the index for sorting by date.

Step2: Calculate the daily rate of return as an input feature of the model. The specific method is to use the logarithmic difference between the closing price of the current day and the closing price of the previous day, and then multiply it by 100 to improve the calculation stability and efficiency. The specific formula is as follows:

$$x_{1t} = ((\log(p_t) - \log(p_{t-1})) \times 100)^2 \tag{7}$$

where p_t represents the closing price of the current day, p_{t-1} represents the closing price of the previous day, and $\log$ represents the natural logarithm.

Step3: Calculate the volatility of daily returns using the exponentially weighted moving average (EWMA) model, which gives more weight to recent data and can more sensitively reflect short-term market fluctuations when smoothing noise in the data and

capturing the latest market changes. Meanwhile, StandardScaler is used to standardize numerical features such as price, volume, and yield variance to eliminate the dimensional influence between different features.

Step4: The pre-processed features include opening price, highest price, lowest price, closing price, and rate of return variance calculated by the EWMA model. These processed feature data will be used as input to the model for training and evaluating the model performance. The data are divided into training set and test set according to the proportion of 70% and 30%. The model is also trained using the training set, and the generalization ability of the model is evaluated with the test set.

3.2 Construction of Graph Model

In this research, a GCN model is constructed. The following part will introduce the construction process of the graph model in detail, including the processing operations of nodes and edges, the processing operations of adjacency matrices and the design of network architecture.

Processing of Nodes and Edges

Each node represents a stock in the GCN in this paper, and the node features include preprocessed data, such as the opening price, highest price, lowest price and closing price of the stock. To deal with these features, they are first standardized. By calculating the mean and standard deviation of each feature, the eigenvalue is converted into a standard score (i.e. Z-score), eliminating the dimensional influence between different features, with the normalized formula:

$$z = \frac{x - \mu}{\sigma} \quad (8)$$

where z is the standard score, x is the feature value, μ is the mean value of the feature, and σ is the standard deviation of the feature.

The edges in the graph network indicate the correlation between stocks. When constructing edges of a graph, these edges are defined based on historical price correlations between stocks. By calculating the correlation coefficient of price changes between each pair of stocks, the co-movement of stock prices is considered. The formula for calculating the correlation coefficient is as follows:

$$\rho_{ij} = \frac{\text{cov}(p_i, p_j)}{\sigma_i \sigma_j} \quad (9)$$

where ρ_{ij} represents the correlation coefficient between the stock i and j , $\text{cov}(p_i, p_j)$ represents the covariance of the changing price of the stock i and j , and σ_i and σ_j represent the standard deviation of the changing price of the stock i and j respectively.

To determine which stock pairs should have edges, this paper sets a correlation threshold, with the logic setting for the threshold as follows:

If the correlation coefficient $\rho_{ij} >$ the threshold, an edge is established; Otherwise, no edges are established.

Adjacency Matrix and Normalization

The adjacency matrix is the core matrix describing the connection relationship between nodes in graph structure data. To make the model learn and update node features more effectively, the following treatments are conducted in the adjacency matrix. Adding loops adds loops to each node on the diagonal of the adjacency matrix, so that nodes can retain their information during the feature transfer process.

Degree matrix normalization: A symmetric normalization method $D^{-1/2}AD^{-1/2}$ is used, where D is the degree matrix (whose diagonal elements represent the degree of each node, that is, the number of connections) and A is an adjacency matrix with loops added. This normalization process helps to adjust the scale of feature transfer between nodes, so that no matter how many nodes are connected, each node can contribute information evenly. In this way, the nodes with more connections can avoid occupying an excessive proportion during the feature propagation, to ensure the model stability and efficiency.

Construction of GCN Model

The constructed GCN model consists of four graph convolution layers, each of which is specially designed to process and update the features of nodes in the graph. First, the GCNConv layer uses the adjacency matrix to transfer the information of neighboring nodes to update the state of each node. Then, a batch normalization layer is added after each GCNConv layer to reduce internal covariate offset, stabilize the distribution of inputs in each layer, and make model training more stable. Then, LeakyReLU is selected as the activation function to avoid the common neuron death in ReLU, to better capture complex nonlinear relationships in stock data. Finally, to prevent the model from overfitting, a dropout regularization layer is added after each activation layer to improve the generalization ability of the model. The output layer of the GCN model is a fully connected layer, which integrates and maps the features of all nodes to an output value, that is, the future rate of return variance. Adam optimizer is used for model training, which can adjust the learning rate adaptively according to the gradient change of each parameter, to improve the training efficiency. To prevent overfitting, an early stop strategy is also adopted.

3.3 Comparative Analysis of Model Prediction Results

To improve the credibility of the model, this paper uses the stock variance data predicted by XGBoost and LSTM for comparative analysis. The stock variance data predicted by GCN are shown in Fig. 1, and the forecast data of XGBoost and LSTM are shown in Figs. 2 and 3.

To compare the predicted performance of three models, the average MSE, RMSE and MAE predicted by each model for 9 stocks are calculated, with results in Table 2.

According to Figs. 1, 2 and 3, GCN is more stable in predicting stock return variance, while LSTM and XGBoost have obvious variance mutations in some stock forecasts. For example, LSTM has a significantly higher curve from day 3 to day 4 when predicting NKLA, while XGBoost has a similar situation when predicting VFS. In contrast, the variance predicted by GCN changes smoothly without sudden large fluctuations, which suggests that GCN has more stable predictions and can effectively capture stock market volatility.

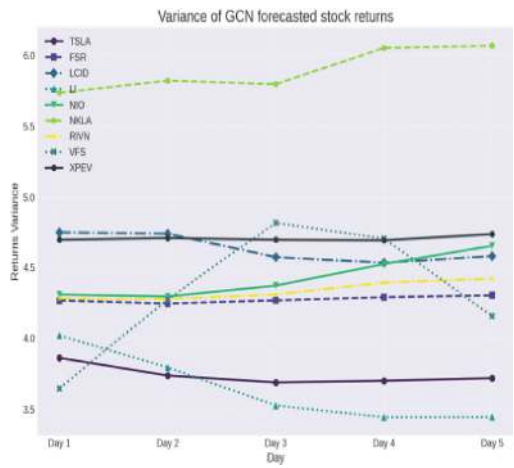


Fig. 1. Return Variance Predicted by GCN (Left)

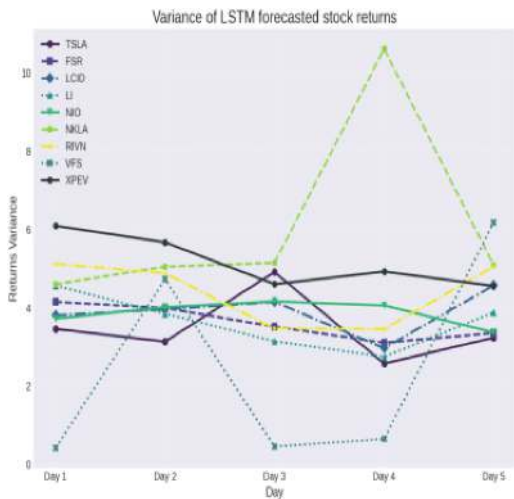


Fig. 2. Return Variance Predicted by LSTM (Middle)

According to Table 2, as for MSE and RMSE, the MSE of the GCN model is 2.934, which is significantly lower than 3.3172 by LSTM and 5.7311 by XGBoost. The RMSE of the GCN model is 1.7469, which is lower than 1.8322 by LSTM and 2.3902 by XGBoost, indicating its excellent performance in overall error control. In terms of MAE, 1.1972 by GCN is very close to 1.1867 by LSTM, both of which are better than 1.4572 by XGBoost, which proves that the GCN model performs well in reducing the deviation between the predicted value and the real value.

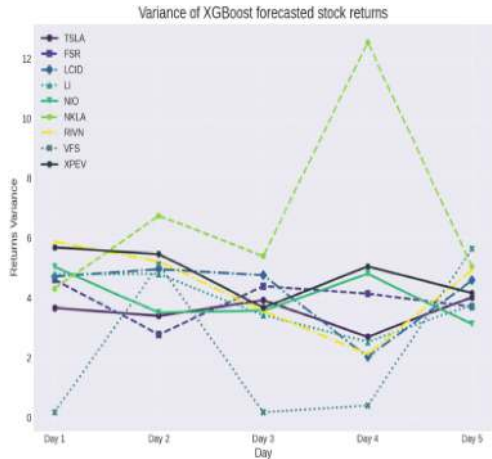


Fig. 3. Return Variance Predicted by XGBoost (Right)

Table 2. Comparison of Prediction Errors of GCN, LSTM and XGBoost

Model	MSE	RMSE	MAE
GCN	2.9340	1.7469	1.1972
LSTM	3.3172	1.8322	1.1867
XGBoost	5.7311	2.3902	1.4572

Comprehensive analysis shows that the prediction of the GCN model is more stable and performs well in multiple error indicators, which proves its effectiveness and accuracy in predicting the variance of new energy vehicle stock returns.

4 Solution of Mean-CVaR Portfolio Model

4.1 Establishment and Solution of Model

In this research, a mean-conditional value-at-risk (CVaR) portfolio model is constructed to solve the optimal portfolio. The flow chart is shown in Fig. 4.

The number of selected stocks n is 9, the confidence α is 0.95, the number of stock returns J is Day5, and the variance is 1. The return variance K of each stock predicted by the GCN model and average return rates of the last week r_k are collected, with these data stored in the return variance matrix and return array respectively. The following equivalent model is obtained:

$$\min_{x, \theta} \hat{F}_\alpha(x, \theta) = \theta + \frac{1}{5(1 - 0.95)} \sum_{k=1}^5 d_k \tag{10}$$

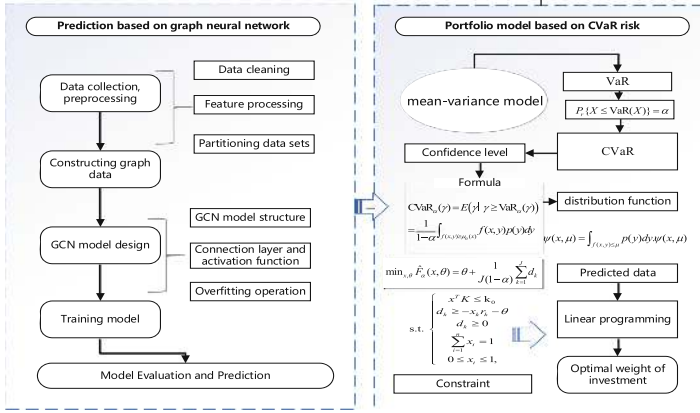


Fig. 4. Mean-CVaR Portfolio Model Based on GNN

$$s.t. \begin{cases} x^T K \leq 1 \\ d_k \geq -x_k r_k - \theta \\ d_k \geq 0 \\ \sum_{i=1}^n x_i = 1 \\ 0 \leq x_i \leq 1, \end{cases} \quad (11)$$

GCN, XGBoost and LSTM models are used to predict the variance of the return rate and the average return rate of the last week, with these data substituted into the mean-CVaR portfolio model to create and solve the optimization problem for each day. Besides, the optimal weight, VaR threshold and risk measure are stored for each day. The cvxpy in Python is to solve the linear programming, and obtain the optimal weight, VaR threshold and weekly average return rate for each three models. The VaR thresholds of three models are in Table 3, the weekly average return rates are in Table 4, and the heat maps based on the optimal weights of three models are in Figs. 5, 6 and 7.

Table 3. VaR Thresholds for Three Models

VaR Threshold	GCN	LSTM	XGBoost
Day1	0.0423	0.2217	0.2385
Day2	0.0332	0.2243	0.2479
Day3	0.0552	0.2131	0.2426
Day4	0.0638	0.1969	0.2523
Day5	0.0425	0.2411	0.2588

Table 4. Weekly Average Return Rates of Three Models

Model	Average Weekly Return Rates
GCN	2.3812
LSTM	1.9769
XGBoost	1.8112

4.2 Comparative Analysis of Models

Firstly, according to Figs. 5, 6 and 7, the weight distribution of the GCN model is more stable. For example, the weight of the FSR is between 27.55% and 28.24% over five days, the weight of the LI steadily increased from 25.90% to 29.76%, and the weight of the NIO is between 16.81% and 18.41%. This shows that the GCN model has good stability in capturing market dynamics, and the weight distribution is uniform, which helps to diversify investment risks. In contrast, the weights of the XGBoost and LSTM models are volatile and concentrated. For example, in the XGBoost model, the weight of FSR is 43.74% on Day1, which suddenly increases to 67.69% on Day2. In the LSTM model, the weight of NKLA suddenly increased to 10.61% on Day4, while it was basically 0 in the remaining days. Finally, it can be seen from the weight distribution of the GCN model that FSR, LI and NIO stocks have higher weights within five days with stability and high expected returns, which should be given priority. TSLA, NKLA, VFS and XPEV stocks have medium weights and need to be adjusted flexibly. LCID and RIVN stocks have low weights, low expected returns and large fluctuations, so they need to be treated with caution.



Fig. 5. GCN Stock Weight Heat Map Figure (Left)



Fig. 6. LSTM Stock Weight Heat Map (Middle)



Fig. 7. XGBoost Stock Weight Heat Map (Right)

Furthermore, according to Table 3, the VaR threshold of the GCN model is significantly lower than that of the LSTM and XGBoost models. Throughout the five-day prediction period, the VaR threshold for the GCN model fluctuated between 0.0332 and 0.0638, while the VaR threshold for LSTM and XGBoost fluctuated between 0.1969 and 0.2588. The lower VaR threshold indicates that the GCN model can better control risk at the same confidence level. Table 4 shows the average weekly return rates of the three models. The average weekly return rates of the GCN model is 2.3812%, which

is significantly higher than 1.9769% by LSTM and 1.8112% by XGBoost. This result proves that the GCN model not only performs well in controlling risks, but also has obvious advantages in improving portfolio returns.

Overall, the mean-CVaR portfolio model based on GCN is better than other models in terms of risk resistance and economic benefits. The GCN model can be chosen to optimize the investment portfolio more stably and improve the reliability of decision-making.

5 Research Conclusions and Prospects

Given the rationality of the financial portfolio model, this paper proposes a mean-CVaR portfolio model based on GNN. The variance of the stock future return is predicted by the GCN and compared with XGBoost and LSTM. The results show that the GCN model has advantages in MSE, RMSE and MAE indicators, and the weight distribution is more stable. In addition, the GCN model has a higher weekly average return rate and a smaller VaR risk threshold. In summary, the method based on GCN is superior to other models in terms of risk resistance and economic benefits.

The mean-CVaR portfolio model based on GNN proposed in this paper provides a new path for financial portfolio optimization. In practical applications, the model still has room for improvement, such as combining more market indicators to improve prediction accuracy, optimizing the model and using stronger computing power to promote computing efficiency. Meanwhile, future research can further optimize the predictive ability of GNN under different market environments and investment strategies in combination with time series features and heterogeneous data characteristics, so as to verify its generalization ability and practical application value.

References

1. Yi, S.: Application of statistics to portfolio optimization. *Highlights in Sci. Eng. Technol.* **88**, 753–759 (2024)
2. Markowitz, H.M.: Portfolio Selection. *J. Financ.* **7**(1), 77 (1952)
3. Jorion, P.: Value at Risk: The New Benchmark for Controlling Market Risk (2000)
4. Basak, S., Shapiro, A.: Value-at-risk-based risk management: optimal policies and asset prices. *The Rev. Fin. Stud.* **14**(2), 371–405 (2001)
5. Yiu, K.: Optimal portfolios under a value-at-risk constraint. *J. Econ. Dyn. Control* **28**(7), 1317–1334 (2004)
6. Nguyen, T.: Selection of the right risk measures for portfolio allocation. *Int. J. Monetary Econ. Fin.* **7**(2), 135 (2014)
7. Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **2**, 270 (2018)
8. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. *ACM* (2016)
9. Wang, Y., Guo, Y.: Application of improved XGBoost model in stock forecasting. *Computer Engineering and Applications* (2019)
10. Sun, Y.: Stock selection strategy based on random forest and support vector machine. *ICSEB 2020:2020 The 4th International Conference on Software and e-Business* (2020)
11. Bedoui, R., Benkraiem, R., Guesmi, K., et al.: Portfolio optimization through hybrid deep learning and genetic algorithms vine Copula-GARCH-EVT-CVaR model. *Technol. Forecast. Soc. Chang.* **197**, 122887 (2023)

12. Rokhsatyazdi, E., Rahnamayan, S., Amirinia, H., et al.: Optimizing LSTM Based Network for Forecasting Stock Market. 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE (2020)
13. Chen, W., Zhang, H., Mehlawat, M.K., et al.: Mean-variance portfolio optimization using machine learning-based stock price prediction. *Appl. Soft Comput.* **100**(1), 106943 (2021)
14. Zhong, X., Enke, D.: Forecasting daily stock market return using dimensionality reduction. *Expert Syst. Appl.* **67**, 126–139 (2017)
15. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. Thirty-second AAAI conference on artificial intelligence (2018)
16. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. *Advances in Neural Info. Proc. Sys.* **30** (2017)
17. Veličković, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. arXiv preprint [arXiv:1710.10903](https://arxiv.org/abs/1710.10903) (2017)
18. Xu, D., Zhu, Y., Choy, C.B., et al.: Scene graph generation by iterative message passing. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5419 (2017)
19. Jie, Z., Ganqu, C., Shengding, H., et al.: Graph neural networks: a review of methods and applications. *AI Open* **1**, 57–81 (2020)
20. Joan, B., Wojciech, Z., Arthur, S., et al.: Spectral networks and locally connected networks on graphs. arXiv preprint [arXiv:1312.6203](https://arxiv.org/abs/1312.6203) (2013)
21. Kipf Thomas, N., Max, W.: Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, pp. 24–26 (2017)



Correlation Analysis of Corporate Culture and Employee Behavior: A Predictive Model Based on Deep Learning

Meichao Song^(✉) and Li Gao

Shandong Business Institute, Yantai, Shandong, China

375663922@qq.com

Abstract. As an important part of the core competitiveness of enterprises, the shaping and influence of corporate culture on employee behavior cannot be ignored. Employee behavior, as the external embodiment of enterprise culture, is directly related to the operational efficiency and market competitiveness of the enterprise. Corporate culture not only covers the values, beliefs, customs and practices of the enterprise, but is also reflected in the daily work behavior and decision-making of employees. A positive and healthy corporate culture can stimulate the spirit of innovation and teamwork among employees, thus improving their job satisfaction and loyalty. This paper delves into the correlation between corporate culture and employee behavior, and proposes a deep learning-based prediction model to quantify and predict this relationship. In this paper, a deep learning-based prediction model is constructed for quantifying and predicting the correlation between corporate culture and employee behavior. The results of the model can not only provide enterprises with profound insights about the relationship between corporate culture and employee behavior, but also provide a reference basis for them to formulate more accurate human resource management strategies. The deep learning-based prediction model provides a new, quantitative analysis tool for enterprises.

Keywords: Deep Learning · Predictive Model · Corporate Culture · Employee Behavior · Correlation Analysis

1 Introduction

As a powerful machine learning method, deep learning has been widely used in various prediction and classification tasks. Deep learning techniques also have great potential in the study of the correlation between corporate culture and employee behavior. By collecting and analyzing a large amount of data on corporate culture and employee behavior, deep learning models are able to automatically extract key features from the data and construct complex prediction models, thus providing a scientific basis for the optimization of corporate culture and the improvement of employee behavior. Applying research results to actual enterprise management and human resource management is the key to increasing the value of research. Deep learning-based prediction models are

applied to corporate culture building and employee behavior management, and their effects and challenges in practical applications are explored.

Scholars generally agree that corporate culture has a significant impact on employee behavior. For example, Schein, E. H proposed three levels of corporate culture: basic assumptions, values, and symbols, and noted that each of these levels affects employee behavior and attitudes [1]. Similarly, Kotter, J. P., and Heskett, J. L. found in their study that companies with strong corporate cultures excel in employee satisfaction, loyalty, and performance [2]. These studies provide a theoretical basis for understanding the relationship between corporate culture and employee behavior.

Regarding the application of deep learning in the prediction of correlation between corporate culture and employee behavior, some scholars have also explored it in recent years. For example, Denison, D. R. et al. used a deep learning model to predict the employee turnover rate, and by analyzing employee behavioral data and corporate culture characteristics, the model was able to accurately predict the employees' intention to leave the company [3, 4]. Schein, E. H used a deep learning model to predict the innovative behavior of employees, and found that the innovative atmosphere and incentive mechanism in the corporate culture had a significant impact on the innovative behavior of employees [5]. These studies provide feasible ideas for deep learning applied to the prediction of correlation between corporate culture and employee behavior. Deal, T. E. et al. builds more complex models to capture the nonlinear relationship between corporate culture and employee behavior, achieving a deeper understanding of the mechanism of interaction between the two [6, 7]. Tsai, W. conducts research that integrates interdisciplinary fusion, synthesizing theories and methods from different disciplines to explore the relationship between corporate culture and employee behavior [8]. Weick, K. E research focuses on building real-time, dynamic predictive models in order to promptly capture these changes and make adjustments accordingly [9].

However, current research still has some shortcomings. First, most of the existing studies use traditional statistical analysis methods, which have limited ability to process and analyze big data, making it difficult to dig deeper into the complex relationships in the data. Second, the definitions and measurements of corporate culture and employee behavior are diverse and lack uniform standards and indicators, leading to questions about the reliability and comparability of research results. Finally, most of the existing studies focus on the direct impact of corporate culture on employee behavior, while ignoring the impact of other potential factors, such as employees' personal characteristics and organizational environment.

In order to solve these problems, this paper proposes a deep learning-based correlation prediction model of corporate culture and employee behavior on the basis of existing scholars' research. By integrating theories and methods from multiple disciplines, the model aims to comprehensively and deeply explore the relationship between corporate culture and employee behavior, and provide scientific guidance for the optimization of corporate culture and the improvement of employee behavior. In this paper, through the integration of the diversity of data sources, the evaluation and optimization of the prediction model, and the use of deep learning technology to explore the correlation between corporate culture and employee behavior in depth, and to provide scientific guidance for the optimization of corporate culture and the improvement of employee behavior.

2 Predictive Model Construction Based on Deep Learning

2.1 Model Building Methodology

The expansion of enterprise business scale exposes them to massive data information and complex business scenarios. In order to effectively understand and predict the correlation between corporate culture and employee behavior, deep learning techniques are used to construct prediction models. Deep learning, with its powerful feature extraction and complex relationship modeling capabilities, provides new solution ideas for predicting the correlation between corporate culture and employee behavior. In constructing the prediction model based on deep learning, this paper mainly adopts neural networks as the basic architecture, especially RNN and LSTM to deal with data with time-series characteristics [10]. Meanwhile, in order to capture the nonlinear relationship between corporate culture and employee behavior, this paper introduces MLP as a feature extraction layer.

2.2 Model Building Process

Firstly, data related to corporate culture and employee behavior need to be collected. These data may come from a variety of sources such as questionnaires, interviews, internal corporate databases, and so on. A series of preprocessing operations need to be performed after the data are collected, including data cleaning, missing value filling, feature coding and so on. In the feature extraction stage, MLP is utilized to nonlinearly transform the corporate culture features in order to extract more representative features. The basic structure of MLP consists of input layer, hidden layer and output layer.

Since corporate culture and employee behavior may have temporal characteristics (e.g., changes in employee behavior over time), RNN or LSTM is used to model the temporal sequence of the data after feature extraction. The basic structure of RNN consists of an input layer, a hidden layer, and an output layer, as well as a cyclic connection. In RNN, the output of the hidden layer depends not only on the current input, but also on the output of the hidden layer at the previous moment. As an improved form of RNN, LSTM could better capture long-term dependencies by introducing a gating mechanism.

The forward propagation process of LSTM involves several mathematical computational steps, where the forgetting gate computation determines the value of what information is discarded from the previous cell state, as shown in Eq. 1 below:

$$F_t = \sigma(Q_f \cdot [y_{t-1}, x_t] + p_f) \quad (1)$$

In which, Q_f is the weight matrix of the forgetting gate, p_f is the bias term, σ is the sigmoid activation function, y_{t-1} is the hidden state of the previous time step, and x_t is the input of the current time step. The input gate calculates the value that determines which new information will be stored in the cell state, as shown in Eq. 2 below:

$$I_t = \sigma(Q_i \cdot [y_{t-1}, x_t] + p_i) \quad (2)$$

In which, Q_i is the weight matrix of the input gate and p_i is the bias term. Meanwhile, the value of the new candidate cell state is calculated as shown in Eq. 3 below:

$$\tilde{C}_t = \tanh(Q_c \cdot [y_{t-1}, x_t] + p_c) \quad (3)$$

In which, Q_c is the weight matrix of candidate cell states, p_c is the bias term, and $\tanh$ is the hyperbolic tangent activation function. Combining the outputs of the forgetting gate and the input gate to update the cell state is calculated as shown in Eq. 4 below:

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \quad (4)$$

In which, $\odot$ denotes the multiplication between elements (the Hadamard product). Output Gate calculates the value that determines what information is output from the cell state, as shown in the following equation:

$$O_t = \sigma(Q_o \cdot [y_{t-1}, x_t] + p_o) \quad (5)$$

In which, Q_o is the weight matrix of the output gate and p_o is the bias term. Based on the output gate and the updated cell state, the hidden state of the current time step is calculated as shown in Eq. 6 below:

$$Y_t = O_t \odot \tanh(C_t) \quad (6)$$

Equation 1–6 describe how the LSTM cell computes a new cell state C_t and hidden state Y_t at a given time step t based on the current input x_t , the hidden state y_{t-1} from the previous time step, and the cell state C_{t-1} . These states are then used in the computation of the next time step.

2.3 Model Training and Optimization

During the model training process, the cross-entropy loss function is used as the optimization objective and gradient descent algorithms (e.g., Adam, RMSprop, etc.) are used to update the model parameters. By continuously iterating the training process, it could make the model gradually approximate the real data distribution, thus improving the accuracy of prediction. After the model training is completed, it needs to evaluate the model to verify its performance. Commonly used evaluation metrics include accuracy, recall, F1 value, and so on. If the model performance is poor, we can try to adjust the network structure, change the activation function, adjust the learning rate and other parameters to optimize the model performance. The deep learning-based model training and optimization process is shown in Fig. 1 below.

Export trained models to a format suitable for deployment (e.g., ONNX, PMML, etc.). Deploy the model in the enterprise management environment and continuously monitor the performance of the model. Iteratively update and optimize the model by collecting feedback from enterprise users and business data to continuously improve the model to meet business requirements and enhance model accuracy.

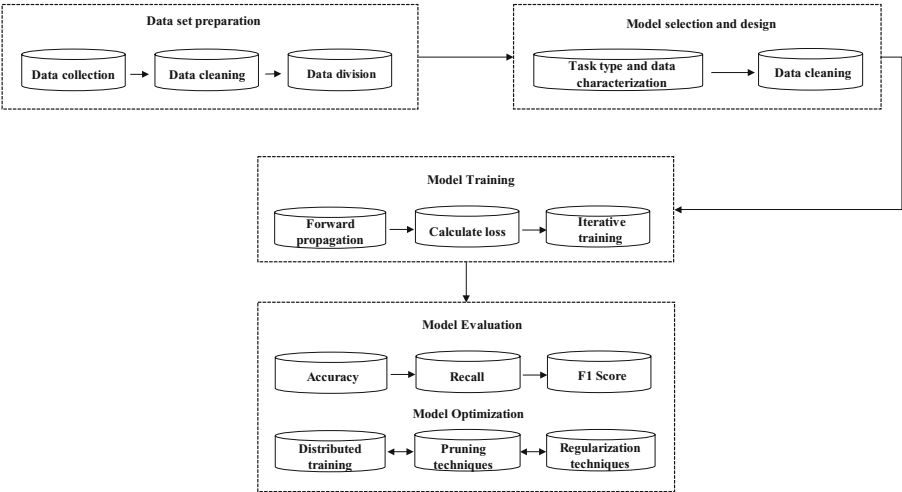


Fig. 1. Process of deep learning-based model training and optimization

3 Correlation Between Corporate Culture and Employee Behavior

3.1 The Impact of Corporate Culture on Employee Behavior

As a guiding and motivating mechanism for the behavior of employees within the enterprise, corporate culture can help employees establish correct work attitudes and values, and stimulate employees' passion for work and sense of responsibility. Through the cultivation of corporate culture, employees can form a common goal and vision with the enterprise, so as to more actively involved in the work. A good corporate culture can create a positive working atmosphere and team spirit, so that employees work more collaboratively and cooperatively. In such an environment, employees can better utilize their professional skills and creativity, and improve work efficiency and quality. At the same time, the concept of teamwork in the corporate culture can also promote communication and collaboration among employees, and enhance the cohesion and centripetal force of the team. The performance of corporate culture influence is shown in Fig. 2 below.

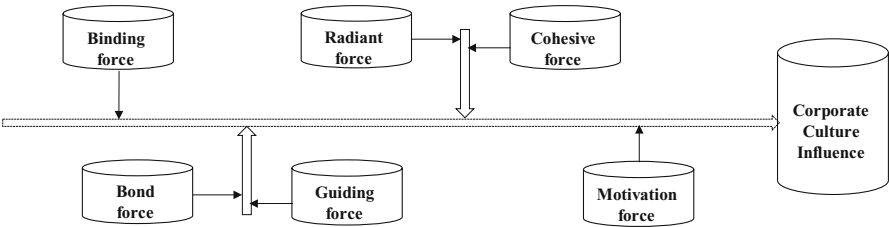


Fig. 2. The performance flow of corporate culture influence

Corporate culture can help employees form a sense of identity and belonging to the enterprise, making them more willing to contribute to the development of the enterprise. When employees identify with the company's culture and values, they will cherish the opportunity to work in the company and strive to create more value for the company. This not only reduces employee turnover, but also attracts more talented people to the organization.

3.2 The Role of Employee Behavior on Corporate Culture

Employees are the inheritors and promoters of enterprise culture. In their daily work, employees pass on the culture and values of the enterprise to the people around them through their behavior and speech, thus influencing more people. At the same time, the behavior of employees can also reflect the actual effect of corporate culture, providing feedback and reference for the development of corporate culture. Employee behavior is a direct reflection of the corporate image and brand. Good employee behavior can shape the good image and brand of the enterprise and enhance the visibility and reputation of the enterprise. On the contrary, bad employee behavior may damage the image and brand of the enterprise and bring negative impact to the enterprise.

Employees are the innovators and promoters of corporate culture. In their daily work, employees may encounter various problems and challenges, and they need to solve problems and meet challenges through their own thinking and efforts. In this process, employees may generate new ideas and creativity, which can provide new ideas and directions for the innovation and development of corporate culture.

3.3 Deep Learning Model-Based Correlation Analysis

By collecting the behavioral data of the employees of an enterprise and the corresponding corporate culture data. Employee behavioral data includes work performance, work attitude, teamwork participation, etc.; corporate culture data includes corporate values, beliefs, norms, etc. These data come from multiple channels, such as the HR system of the enterprise, employee self-assessment, and supervisor evaluation. After collecting the raw data, we pre-processed the data. This includes data cleaning (removing duplicated, missing, and abnormal data), data coding (converting textual data into computer-recognizable values), and data standardization (converting data with different scales into a uniform scale for subsequent analysis). From the preprocessed data, we extracted key features. Corporate culture features include teamwork, innovation, and customer orientation; employee behavioral features include job performance, work attitude, and turnover rate. The selection of these features is based on previous research and theories, as well as an in-depth understanding of the actual situation of the enterprise.

Using a deep learning framework, a prediction model containing a multi-layer neural network is constructed. The model takes corporate culture features as input and employee behavior features as output. By training the model, we expect to find the mapping relationship between corporate culture and employee behavior. During the model training process, back propagation algorithm and gradient descent method are used to optimize the parameters of the model. Techniques such as regularization and dropout are used to

prevent overfitting. The model parameters are continuously iterated and adjusted so as to obtain a model with good predictive performance.

4 Predictive Modeling Application Analysis Results

4.1 Analysis Data of Research Samples

In order to deeply analyze the correlation between corporate culture and employee behavior, this paper carries out data collection and analysis on the employees of an enterprise, as shown below.

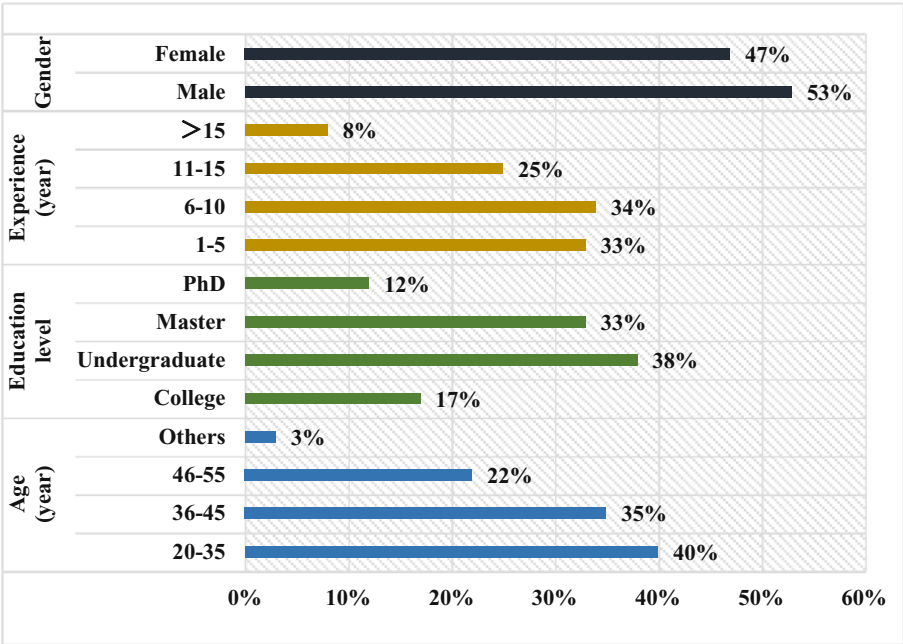


Fig. 3. The typical data of research samples

As can be seen from Fig. 3, the data composition of the selected employee sample elements (education, gender, years of work, age) is shown in Fig. 3 below. As can be seen from the data in the figure, most of the study samples are well educated and most of their working years are concentrated within 10 years.

4.2 Corporate Culture Perception Data Mining

Using the constructed deep learning model, the data on the recognition of corporate culture of the selected samples of corporate employees are analyzed and mined, and the results obtained are shown in Fig. 4 below.

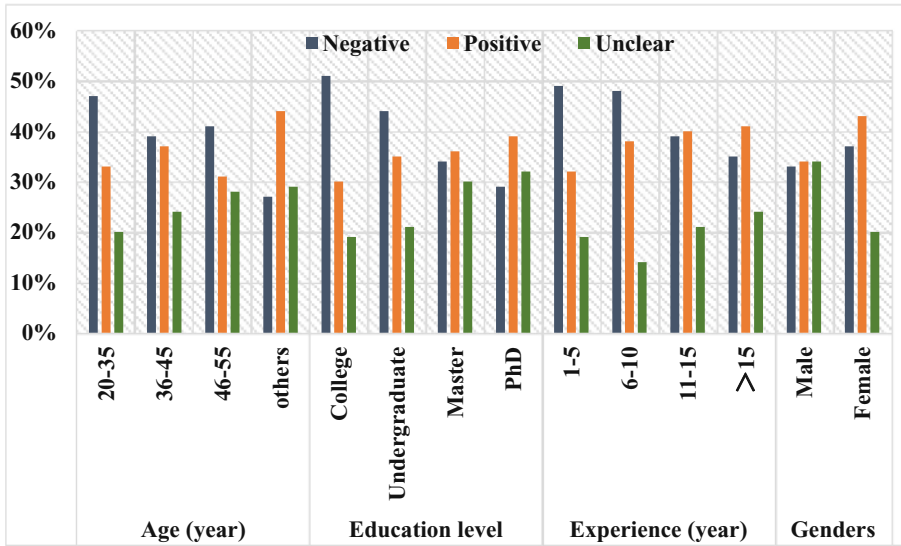


Fig. 4. The data of employees' corporate culture awareness

As can be seen from Fig. 4, with the growth of the working years of enterprise employees, their recognition of enterprise culture also shows an upward trend. Secondly, the education level of employees is positively related to their corporate culture recognition. In addition, the age of the employees and the degree of cognition also shows a positive relationship, and the female employees' cultural cognition is slightly stronger than that of male employees. This analysis further verifies the survivor bias effect, because employees who do not recognize the corporate culture will gradually leave the company, which is why the longer the employees work and the older they are, their recognition of corporate culture will gradually increase.

4.3 Correlation Analysis Between Corporate Culture and Employee Behavior

The correlation analysis of corporate culture and employee behavior was carried out using a deep learning-based prediction model, in which the results of the influence of corporate culture on employee behavior are shown in Table 1 below, in which the positive influence indicates the degree of positive influence of corporate culture on employee behavior, expressed as a percentage. Negative impact indicates the degree of potential or actual negative impact of corporate culture on employee behavior, expressed as a percentage.

As can be seen from the data in Table 1, for the corporate culture that emphasizes innovation, the majority of employees say that their willingness to innovate increases, and only a small number of employees feel that the pressure increases. Under the influence of the corporate culture emphasizing teamwork, the vast majority of employees indicate that teamwork ability increases, and a very small number of employees feel that their individual contributions are neglected. For a corporate culture that emphasizes customer satisfaction, the vast majority of employees provide customer-centered services, and

Table 1. The impact of corporate culture on employee behavior

Corporate Culture	Positive Impact	Negative Impact	No Impact
1. Emphasizing innovation	75%	10%	15%
2. Teamwork	80%	5%	15%
3. Customer satisfaction	90%	3%	7%
4. Fair competition	85%	7%	8%
5. Employee Benefits	95%	2%	3%
6. Leadership style	70%	15%	5%
7. Career development opportunities	85%	10%	5%
8. Ethics	98%	1%	1%

a very small number of employees show indifference in their services. For a culture that emphasizes fair competition, most employees believe the competitive environment within the company is fair. For a corporate culture that emphasizes employee benefits, the vast majority of employees have increased their level of satisfaction. Additionally, for corporate cultures that emphasize career development opportunities and ethics, the vast majority of employees are able to see career paths and abide by the company’s code of ethics. Only for corporate cultures that emphasize leadership style, one-third of employees have reservations or no feelings about leadership style. The data analyzed above shows that a positive corporate culture will have a positive impact on employee behavior to a large extent.

Furthermore, the data in Table 2 below gives data on the counterproductive effects of employee behavior on corporate culture as mined by the deep learning-based predictive model.

Table 2. The counterproductive effects of employee behavior on corporate culture

Employee Behavior	Positive Impact	Negative Impact	No Impact
1. Active participation	85%	10%	5%
2. Practicing values	90%	5%	5%
3. Teamwork	80%	14%	6%
4. Innovation	75%	10%	15%
5. Customer satisfaction	95%	3%	2%
6. Compliance with norms	98%	2%	0%
7. Participation in cultural activities	80%	15%	5%
8. Feedback and suggestions	70%	20%	10%

The data in Table 2 demonstrates how employee behavior affects the shaping, dissemination, and maintenance of corporate culture. Positive impact on corporate culture

indicates how employee behavior contributes to the shaping, dissemination, and maintenance of corporate culture, expressed as a percentage. Similarly, potential challenges to corporate culture indicates the problems or challenges that employee behavior may pose, expressed as a percentage. From the data in Table 2, it can be seen that positive employee behavioral traits also contribute to a great extent to the improvement and adjustment of corporate culture and vice versa.

4.4 Application of Correlation Analysis Results

Based on the results of the correlation analysis between corporate culture and employee behavior, companies can identify which elements of corporate culture have a significant impact on positive employee behavior. Based on these findings, companies can adjust their hiring strategies to find candidates who are more aligned with the values of the corporate culture. For existing employees, companies can develop more targeted training and development programs to help them better understand the culture and incorporate its values into their daily work. Relevance analysis can help companies identify the extent to which employee behaviors match the company's culture, so they can design more accurate incentives. For example, additional rewards and recognition can be given to employees who demonstrate a high degree of teamwork. By rewarding behaviors that are in line with the company's cultural values, the company can further consolidate and reinforce its cultural values and increase employees' sense of belonging and satisfaction.

In addition, if the results of the analysis show that certain elements of corporate culture are negatively affecting employee behavior, companies can consider culture shaping or change. Companies can use the results of the correlation analysis to communicate with employees, explain how corporate culture affects their behavior, and encourage employees to actively participate in the construction of corporate culture. Correlation analysis of corporate culture and employee behavior is an ongoing process. Enterprises need to collect and analyze data on a regular basis to understand whether the impact of corporate culture on employee behavior has changed. Based on the results of the analysis, companies can make timely adjustments to their management strategies and culture-shaping programs to maintain the vitality and effectiveness of their corporate culture.

5 Conclusions

Corporate culture is not only an abstract idea or slogan, but also permeates the daily operations, management decisions and employee interactions of an organization. Through the analysis of deep learning models, this paper finds that values, vision, mission and other factors in corporate culture can significantly predict employees' work attitude, teamwork and innovation ability. Similarly, the feedback effect of employee behavior on corporate culture cannot be ignored. Employees are not only the receivers of corporate culture, but also the transmitters and shapers of corporate culture. Through the model prediction, we find that the employee's behavioral pattern, work attitude and innovation willingness and other factors can in turn affect the formation and development of corporate culture. In the process of enterprise culture construction, it should pay attention to the participation and feedback of employees to form a benign interactive cultural ecology. In summary,

there is a close correlation between corporate culture and employee behavior, and this correlation can be accurately analyzed and predicted by the prediction model based on deep learning. In the future practice of enterprise management, more attention should be paid to the construction and management of enterprise culture, and at the same time, the feedback effect of employee behavior on enterprise culture should be paid attention to, in order to realize the common development of the enterprise and employees.

Acknowledgements. Yantai Zhuoming Property Management Co., Ltd. Corporate Culture Core Concept Construction Plan (Shandong Business Institute Horizontal Research Project, No. 2024HXZX013).

References

1. Schein, E.H.: *Organizational Culture and Leadership*, 5th ed., pp. 147–150. Jossey-Bass (2010)
2. Kotter, J.P., Heskett, J.L.: *Corporate culture and performance*. Free. Press. **5**(2), 227–295 (2005)
3. Denison, D.R., Mishra, A.K.: Toward a theory of organizational culture and effectiveness. *Organ. Sci.* **6**(2), 104–123 (2005)
4. O'Reilly, C.A., Chatman, J.: Organizational culture as a strong context for personnel psychology. *J. Organ. Behav.* **12**(2), 77–88 (2012)
5. Schein, E.H.: How can organizations learn faster? the challenge of entering the green room. *Sloan Management Review* **37**(2), 7–20 (2016)
6. Deal, T.E., Kennedy, A.A. *Corporate Cultures: The Rites and Rituals of Corporate Life*. Addison-Wesley Publishing Company, pp. 131–148 (2002)
7. Cameron, K.S., Quinn, R.E.: *Diagnosing and changing organizational culture: based on the competing values framework*. Addison-Wesley Publishing Company. **5**(1), 119–126 (2008)
8. Tsai, W., Ghoshal, S.: Social capital and value creation: the role of intrafirm networks. *Acad. Manag. J.* **41**(4), 464–476 (2018)
9. Weick, K.E.: *Sensemaking in organizations*. Thousand Oaks, CA: Sage Publications **4**(2), 107–116 (2009)
10. Kim, S., Lee, H.: The Impact of Organizational Culture and Leadership Style on Job Satisfaction and Organizational Commitment. *Leadersh. Org. Dev. J.* **37**(7), 930–950 (2016)



Big Data Mining Techniques and Their Optimisation in Financial Data Analysis

Qingle Chen^(✉)

Shandong University of Finance and Economics (Shengjing Campus), Jinan, Shandong, China
cq1cq120232023@qq.com

Abstract. Big data mining techniques are crucial in modern financial data analysis. Large amounts of data are analysed in depth through advanced algorithms to extract the required information and transform and mine it appropriately to gain valuable insights. Big data mining objects include structured, semi-structured and unstructured data, and its core process covers data selection, preprocessing, transformation, mining and model evaluation, with each step providing a theoretical basis for systematic analysis. Stochastic decision trees are particularly important in financial data classification, which can increase the generalisation ability and robustness of the model, avoid overfitting and improve the prediction ability. Combining the methods of discriminant matrix and artificial weights can optimise the decision tree structure and improve the classification accuracy. Experiments show that the classification accuracy of the stochastic decision tree is relatively stable in different risk categories, but the accuracy of the high-risk category is lower, mainly due to the insufficient amount of high-risk data. By adding 500 high-risk data and adopting the stratified sampling method, the classification accuracy of the high-risk category was significantly improved while maintaining the high accuracy of other categories. Overall, the stratified sampling method significantly improves the classification model performance and provides reliable technical support for financial institutions' risk assessment. The successful application of big data mining techniques in financial data analysis demonstrates its great potential in handling complex financial data and enhancing risk management capabilities.

Keywords: Big Data Mining · Financial Data Analysis · Stochastic Decision Tree · Data Classification · Stratified Sampling

1 Introduction

As an important part of knowledge discovery, big data mining technology uses advanced computer algorithms to analyse large amounts of data in depth, extract the required information from huge databases, and transform, mine, and exploit these data appropriately to gain valuable insights. First, data selection is used to filter out the data sets that meet the requirements from the data sources; then, data preprocessing is performed to clean and organise the data to remove noise and outliers; subsequently, the preprocessed data is converted into a format that facilitates the subsequent mining analysis; then, data mining techniques are applied to conduct in-depth analyses of the converted data to excavate

the potential patterns and relationships; finally, the mined models are evaluated and validated to ensure their effectiveness and accuracy [1].

By including a stochastic decision tree scheme into the model, the classification issue of financial data in financial data analysis may be successfully resolved. In particular, the relevance of each characteristic under various mining aims differs in financial data warehouses, thus while building the decision tree, the significance of each feature must be statistically analyzed. Presently, information entropy-based schemes and the discriminant matrix serve as the foundation for most widely used approaches for assessing the significance of characteristics [2]. But using the discriminant matrix alone is not enough to accurately capture the true significance of features since financial data is extremely specialized. Therefore, this project adjusts and intervenes in the discriminant matrix by introducing artificial weights to improve the accuracy of attribute weight calculation. In addition, the stochastic decision tree scheme increases the generalisation ability and robustness of the model by introducing randomness [3]. The combination of the discriminant matrix and artificial weights can reflect the importance of attributes more accurately and optimise the structure of the decision tree, thus improving the effectiveness and accuracy of financial data classification.

The validation data used in this study are derived from the financial statistics of more than 2,000 companies cooperating with B commercial banks in country A, and the time period is uniformly from 2019 to 2022. Since the attribute classification of financial information data sheets is based on the banks' transaction databases, the financial information data sheets provided by these banks can be converted into 16 attributes that can clearly show the financial status of companies. These attributes not only cover key financial indicators of profitability, solvency and operational efficiency, but also provide a comprehensive picture of a company's financial health and operating results [4]. These attributes are selected based on their importance and commonality in financial analyses and can provide valuable insights to financial institutions and researchers to support their decision-making process. Detailed analyses of these attributes provide insights into a company's financial condition, which in turn provides a scientific basis for risk assessment and customer management in commercial banks.

To further improve the classification accuracy, the researchers added an additional 500 high-risk levels to the training dataset. To ensure the adequacy of the training data for the high-risk level, the study used stratified sampling instead of the original random sampling method, where the initial data were stratified according to small, medium-low, medium-high and high-risk, and then random sampling was conducted within each stratum [5]. With this stratified random sampling method, the researchers succeeded in increasing the classification accuracy of the high risk category while maintaining high classification accuracy of the other risk categories.

In summary, this study effectively solved many problems in financial data classification by introducing big data mining technology and random decision tree scheme combined with stratified sampling method. The results show that this method not only improves the accuracy and stability of the classification model, but also provides more accurate and reliable technical support for banks and other financial institutions in risk assessment and management [6]. The study provides valuable experience and insights

for further optimising the classification algorithm in the future, and also lays a solid foundation for the application and development of financial data mining technology.

2 Related Work

In the contemporary financial sector, big data mining-based financial data analysis and applications play a significant role. Big data mining technology provides in-depth analysis of massive data through advanced computer algorithms, extracts valuable information from huge databases, and appropriately transforms, mines, and exploits these data to gain valuable insights. This technique is mainly applied to deal with structured data, semi-structured data, and unstructured data, and its core processes include data selection, data preprocessing, data conversion, data mining, and model evaluation [7].

Firstly, in the data selection stage, the foundation for subsequent analysis is laid by screening data sets from data sources that meet the requirements. Next, the data preprocessing session cleans and organises the data to remove noise and outliers to ensure the quality and consistency of the data. Subsequently, the preprocessed data needs to be formatted to make it suitable for subsequent mining analysis. In the data mining stage, data mining techniques are applied to analyse the converted data in depth and mine out potential patterns and relationships. Finally, the mined models are evaluated and validated through model evaluation to ensure their validity and accuracy. The complete flowchart shows each step from data source selection to final model evaluation, and each step is crucial, providing a systematic approach and scientific theoretical basis for big data analysis [8].

The significance of characteristics in financial data warehouses varies depending on the mining aims. Thus, while building the decision tree, the significance of each characteristic must be quantitatively examined. Discriminant matrix-based and information entropy-based systems are two popular approaches. In this study, the discriminant matrix scheme is used to assess the importance of attributes, and the discriminant matrix is adjusted and intervened by introducing artificial weights to improve the accuracy of attribute weight calculation [9]. The stochastic decision tree scheme increases the generalisation ability and robustness of the model by introducing randomness, avoids the overfitting problem, and improves the model's ability to predict new data.

The validation data used in the experimental analysis section are derived from the financial statistics of more than 2,000 companies cooperating with Commercial Bank B in Country A. The time period of the validation data is from 2019 to 2022. These data were converted into 16 attributes based on the bank's transactional database that provide a clear picture of a company's financial position. The attributes cover key financial indicators such as a company's profitability, solvency and operational efficiency, providing a comprehensive picture of a company's financial health and operating results. By analysing these attributes in detail, it can provide financial institutions and researchers with valuable insights to support their decision-making process.

In the practical application, based on the data from 2019 to 2022, the company's risk is classified into four categories: high risk, medium-high risk, medium-low risk, and low risk. The results of the study show that constructing 10 random decision trees for decision analysis works best. In the experiment, the stratified sampling method significantly

improved the classification accuracy of the high risk category, while the other risk categories maintained a high classification accuracy. Specifically, the average classification accuracy was 90.76% for the small-risk category, 88.02% for the medium-low-risk category, 76.56% for the medium-high-risk category, and 73.26% for the high-risk category, which was higher than the previous 53.94%.

Overall, the use of stratified sampling method not only performs well in the classification of high-risk categories, but also improves the accuracy and stability of the classification model as a whole. Through reasonable data sampling and balancing, the distribution of data across categories significantly improves the performance of the classification model [10]. This provides banks and other financial institutions with more accurate and reliable technical support in risk assessment and management, as well as valuable experience and insights for further optimising the classification algorithm.

3 Modelling

Big data mining technology is an important part of knowledge discovery, which uses advanced computer algorithms to analyse large amounts of data in depth, extract the required information from huge databases, and transform, mine, and utilise this data appropriately to gain valuable insights. The objects of big data mining mainly include structured data, semi-structured data, and unstructured data. Its core processes usually include data selection, data preprocessing, data transformation, data mining and model evaluation [11]. First, the data sets that meet the requirements are filtered from the data sources through data selection; then, data preprocessing is performed to clean and organise the data to remove noise and outliers; subsequently, the preprocessed data is converted into a format that facilitates the subsequent mining analysis; then, the converted data is analysed in depth by applying the data mining techniques to unearth the potential patterns and relationships; finally, the mined models are evaluated and validated to ensure their effectiveness and accuracy. Figure 1 shows this complete flowchart, from the selection of data sources to the evaluation of the final model, each step is crucial and closely related, providing a systematic approach and scientific theoretical basis for big data analysis.

The classification problem of financial data can be effectively solved by introducing a stochastic decision tree scheme in the model, as shown in Fig. 2. The figure shows the specific process of case deletion. First, obtain the current time T_n , and then judge whether the input time In is greater than T . If so, continue to obtain the CM value; otherwise, the process ends. Next, determine whether CN is greater than CM , if yes, calculate the minimum difference value D in the case base and determine whether D is greater than the threshold Y_n . If greater, delete the relevant case; if not, select the case with the least number of matches from the case base and delete it.

The stochastic decision tree scheme is able to increase the generalisation ability and robustness of the model by introducing randomness. During the decision tree construction process, the overfitting problem is avoided by randomly selecting attributes and splitting points to improve the model's prediction ability for new data. Combining the methods of discriminant matrix and artificial weights can reflect the importance of attributes more accurately and optimise the structure of the decision tree, so as to improve the effect and accuracy of financial data classification [12].

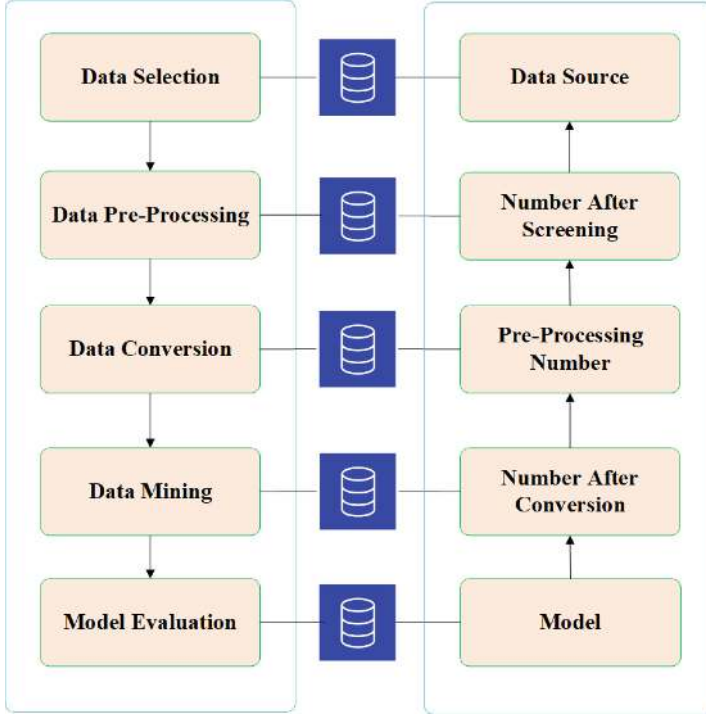


Fig. 1. Flow of big data mining

Through the above methods, not only can the classification problem of financial data be effectively solved, but also the overall performance of the model can be improved to ensure that it can still maintain high prediction accuracy and stability in the face of a complex and changing financial environment. The case deletion process shown in Fig. 2 is an important part of achieving this goal, optimising the data processing process and improving the efficiency and effectiveness of data mining through reasonable case management and deletion strategies.

Diagonal matrix of $|u| \times |u|$. These are the definitions of each term:

$$C_{ij} = \begin{cases} \{\alpha \in A | \alpha(x_i) \neq \alpha(x_j)\} & d(x_i) \neq d(x_j), d(x) \in D \\ \phi(x_i) = d(x_j), d(x) \in D & \end{cases} \quad (1)$$

Set up every a_i in A such that $w(a_i) = 0$. Calculate each term of the diagonal matrix in the C_{jk} resolution matrix.

$$w(a_i) = w(a_i) + |C_{jk}|, a_i \in C_{jk}, 0 < k < j = |U| \quad (2)$$

The basis of all attributes in the equation above is $|A|$, and the base of the discrimination matrix C_{jk} is $|C_{jk}|$. $W_{ai} = w(a_i) + w^I(a_i)$ is the weight of attribute a_i . If you want to increase the weight of a_i , set $w^I(a_i)$ to a positive value, and if you want to decrease

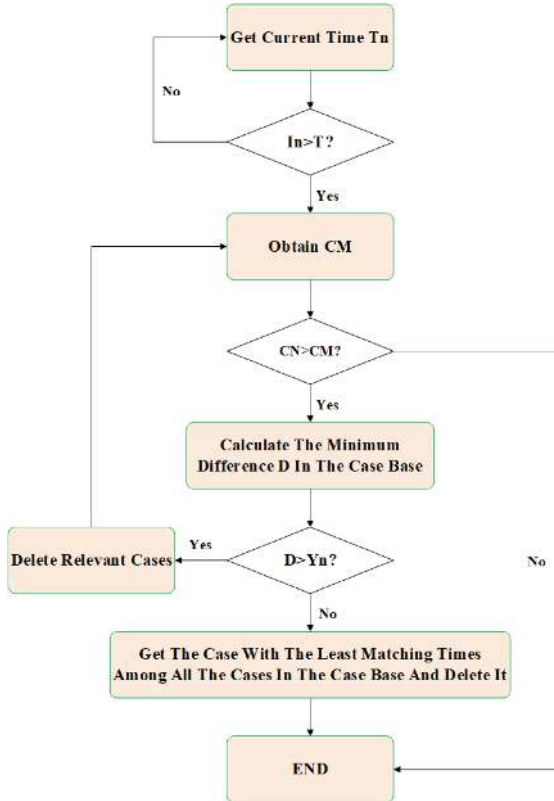


Fig. 2. Case deletion process

the weight of a_i , set it to a negative value. The system allows for manual weight correction after it presents the weights. To do this, you must add the correction coefficients $w^I(a_i)$, $-1 \leq w^I(a_i) \leq 1$.

A method for assessing attribute weights in complex systems by means of a discriminant matrix. Firstly, the initial weights are calculated based on the frequency of attributes in the discriminant matrix, and then the initial weights are allowed to be manually adjusted through the introduction of correction coefficients to more accurately reflect the importance of the attributes in practical applications. The setting range of the correction coefficient is $[-1, 1]$, which ensures the flexibility and moderation of the adjustment. After the system presents the initial weights of each attribute, users can increase or decrease the weights of certain attributes according to their specific needs in order to better adapt to the actual situation and data characteristics. This approach not only improves the flexibility of the system, but also enhances its ability to adapt to complex and changing environments, making the data mining and classification process more scientific, accurate and efficient.

4 Experimental Analysis

The financial data of over 2,000 businesses in nation A that collaborate with B commercial banks provided the validation data for this research. The validation data spans a single time period from 2019 to 2022. Since the attribute division of the data sheet is based on the bank's transaction database, as shown in Table 1, the financial information data sheet provided by these banks may be turned into 16 attributes that may accurately describe the financial status of the firm.

Table 1. Data sheet on financial information

Attribute	Code Attribute	Attribute	Code Attribute
G1	Return on assets	G9	Inventory turnover ratio
G2	Gearing ratio	G10	Interest cover multiplier
G3	Net profit on total assets	G11	Total assets turnover ratio
G4	Return on assets	G12	Accounts receivable turnover ratio
G5	Operating income net profit ratio	G13	Fixed assets turnover ratio
G6	Quick ratio	G14	Accounts receivable turnover ratio
G7	Current ratio	G15	Capital adequacy ratio
G8	Fixed assets ratio	G16	Inventory current liability ratio

The above attributes not only cover key financial indicators of a company's profitability, solvency and operational efficiency, but also provide a comprehensive picture of a company's financial health and operating results. These attributes are selected based on their importance and commonality in financial analysis and can provide valuable insights for financial institutions and researchers to support their decision-making process. Detailed analyses of these attributes provide insights into the company's financial condition, which in turn provide a scientific basis for risk assessment and customer management in commercial banks.

Financial analysts categorize the company's risk into four groups based on its real situation in 2023 and associated indicators: high risk, medium-high risk, medium-low risk, and low risk. According to this framework of classification, companies classified as high-risk are those that could go bankrupt between 2021 and 2023; companies classified as medium-high-risk are those that could default; companies classified as medium-low-risk will not default but their financial situation will worsen; and companies classified as low-risk are those that are in sound financial standing and won't default. The study's findings indicate that the most effective technique for choice analysis is to create ten random decision trees. Thus, using a random technique, ten random decision trees were built from the analyzed data in this research, and five trials were carried out to ensure the stability of the decision trees.

The remaining 500 data were utilized as test data to assess the decision trees' categorization performance. Using the training data, the researchers created random decision

trees, which they then cross-checked with the test data to see how well they classified the data. Figure 3 displays the experimental data and illustrates the classification accuracy of many stochastic decision trees under various risk categories:

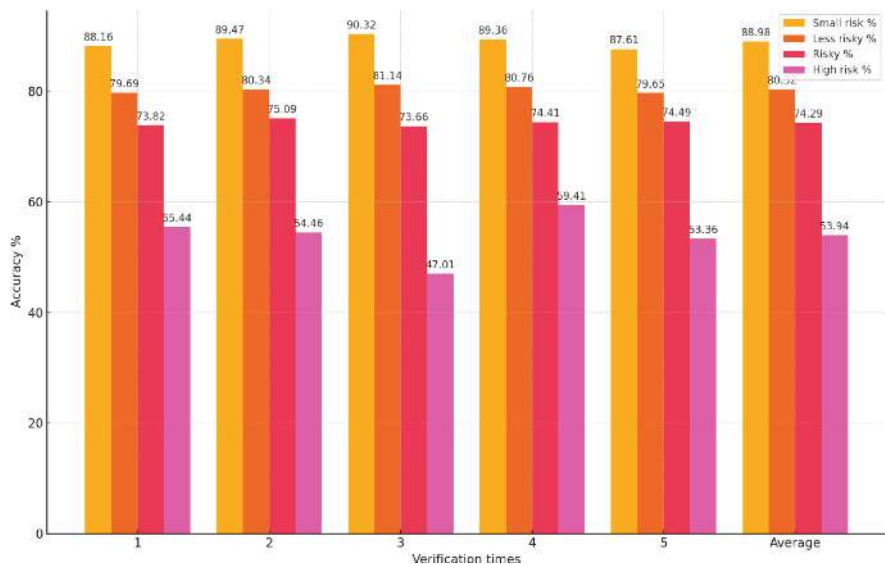


Fig. 3. Comparison of correct classification rates for multiple stochastic decisions

The experimental results show the classification accuracy of multiple stochastic decision trees under different risk categories, from which it can be seen that the algorithm performs relatively consistently in the small, medium-low, and medium-high risk categories, but has significantly lower classification accuracy in the high-risk category. Specifically, the classification accuracies of the five validations ranged from 87.61% to 90.32%, with an average of 88.98% for small risks, from 79.65% to 81.14%, with an average of 80.32% for medium-low risks, and from 73.66% to 75.09%, with an average of 74.29% for medium-high risks. In contrast, the classification accuracy rate for high risk fluctuated considerably, ranging from 47.01% to 59.41%, with an average of only 53.94%.

This phenomenon is mainly attributed to the low amount of high-risk data in the training set, leading to under-trained algorithms in dealing with high-risk categories, which affects the classification accuracy. Although overall, the stochastic decision tree performs well in classifying small and low to medium risks, the low accuracy of the high-risk category suggests the need for further optimisation. Data from high-risk companies tends to have greater uncertainty and complexity, and insufficient samples make it difficult for the model to capture high-risk features, which in turn leads to poor classification performance.

To further improve the classification accuracy, an additional 500 high-risk levels were added to the training dataset to address the problem of insufficient training at high-risk levels. To ensure the adequacy of the high-risk level training data, the study adopted

stratified sampling instead of the original random sampling method. In particular, the original data were first divided into four risk categories: tiny, medium-low, medium-high, and high. Within each of these categories, random sampling was then carried out. Figure 4 displays the categorization outcomes produced by this stratified random sampling technique:

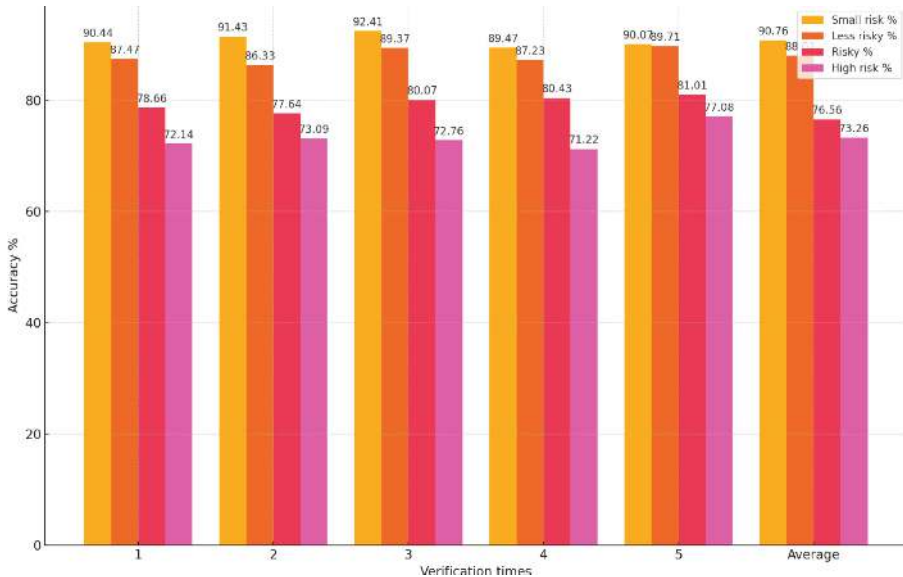


Fig. 4. Comparison of correctness of stratified sampling with multiple random decision trees

From the experimental results, it can be seen that the classification accuracy of multiple random decision trees constructed using stratified sampling method under different risk categories. An in-depth analysis of the data reveals that stratified sampling significantly improves the classification accuracy of high risk categories while maintaining high classification accuracy of other risk categories. Specifically, the classification accuracy of the small risk category ranged from 89.47% to 92.41%, with an average of 90.76%; that of the medium-low risk category ranged from 86.33% to 89.71%, with an average of 88.02%; that of the medium-high risk category ranged from 77.64% to 81.01%, with an average of 76.56%; and that of the high-risk category accuracy ranged from 71.22% to 77.08% with an average of 73.26%.

Compared to the random sampling method, the stratified sampling method significantly improved the classification accuracy of the high-risk categories by almost 20 percentage points from the previous average of 53.94% to 73.26%. This improvement is attributed to the addition of 500 high-risk data in the training dataset and ensuring a balanced distribution of data for each risk category through stratified sampling, which allows the model to fully learn the features of each category during training. This improvement not only solves the problem of insufficient samples of high-risk data, but also reduces the bias of the model on high-risk categories, which significantly enhances the robustness and reliability of the model.

In addition, the data in Fig. 4 also show that stratified sampling also improves the classification accuracy for the medium-high and medium-low risk categories, albeit by a relatively small amount, which still demonstrates the effectiveness of stratified sampling in balancing the distribution of the data and improving the classification accuracy. Especially in the high-risk category, the classification accuracy rate improved from the initial fluctuating range of 47.01%–59.41% to 71.22%–77.08%, demonstrating a more stable classification performance.

Overall, the use of stratified sampling method not only performs well in the classification of high-risk categories, but also improves the accuracy and stability of the classification model in general. This suggests that the distribution of data across categories can significantly improve the performance of classification models through reasonable data sampling and equalisation. The results of this study provide banks and other financial institutions with more accurate and reliable technical support in risk assessment and management, as well as valuable experience and insights for further optimisation of classification algorithms.

5 Conclusion

Financial data analysis and application based on big data mining technology has an irreplaceable and important role in the modern financial industry. Big data mining technology analyses massive data in depth through advanced computer algorithms, extracts valuable information from huge databases, and transforms, mines and exploits these data appropriately to gain valuable insights. This technology is largely employed to deal with structured data, semi-structured data, and unstructured data, and its fundamental process comprises data selection, data preprocessing, data transformation, data mining, and model assessment. Each step of this complete process, from the selection of data sources to the evaluation of the final model, is crucial and provides a systematic approach and scientific theoretical basis for big data analysis.

In the classification problem of financial data, the introduction of stochastic decision tree scheme is particularly important. The classification problem of financial data can be effectively solved by stochastic decision tree, especially when dealing with complex financial data, through which the generalisation ability and robustness of the model can be increased, the overfitting problem can be avoided, and the prediction ability of the model for new data can be improved. This study combines the methods of discriminant matrix and artificial weights, which can reflect the importance of attributes more accurately and optimise the structure of the decision tree, so as to improve the effect and accuracy of financial data classification. The experimental results show that the classification accuracy of multiple stochastic decision trees under different risk categories is relatively stable, especially in the classification of small risk and medium-low risk, but the classification accuracy of high-risk categories is relatively low. This is mainly attributed to the small amount of high-risk data in the training set, which leads to insufficient training of the algorithm in dealing with high-risk categories, thus affecting the classification accuracy.

In order to further improve the classification accuracy, this study added an additional 500 high-risk classes in the training dataset and used stratified sampling instead

of the original random sampling method. The experimental results show that stratified sampling significantly improves the classification accuracy of high-risk categories while maintaining high classification accuracy of other risk categories. Specifically, the classification accuracy of small risk categories averaged 90.76%; medium-low risk categories averaged 88.02%; medium-high risk categories averaged 76.56%; while the classification accuracy of high-risk categories improved to 73.26% from the previous 53.94%. This improvement is attributed to the addition of high-risk data in the training dataset, and stratified sampling to ensure a balanced distribution of data for each risk category, thus allowing the model to fully learn the features of each category during training. This improvement not only solves the problem of insufficient samples of high-risk data, but also reduces the bias of the model on high-risk categories, which significantly enhances the robustness and reliability of the model.

Overall, the use of stratified sampling method not only performs well in the classification of high-risk categories, but also improves the accuracy and stability of the classification model in general. This suggests that the distribution of data across categories can significantly improve the performance of classification models through reasonable data sampling and equalisation. The results of this study provide more accurate and reliable technical support for banks and other financial institutions in risk assessment and management, as well as valuable experience and insights for further optimisation of classification algorithms. The successful practice of big data mining technology in financial data analysis and application fully demonstrates its great potential and broad prospect in handling complex financial data and enhancing risk management capability.

References

1. Hasan, M.M., Popp, J., Oláh, J.: Current landscape and influence of big data on finance. *Journal of Big Data* **7**(1), 21 (2020)
2. Ageed, Z.S., et al.: Comprehensive survey of big data mining approaches in cloud systems. *Qubahan Academic Journal* **1**(2), 29–38 (2021)
3. Qi, C.C.: Big data management in the mining industry. *Int. J. Miner. Metall. Mater.* **27**(2), 131–139 (2020)
4. Kushwaha, A.K., Kar, A.K., Dwivedi, Y.K.: Applications of big data in emerging management disciplines: A literature review using text mining. *Int. J. Info. Manage. Data Insights* **1**(2), 100017 (2021)
5. Haoxiang, W., Smys, S.: Big data analysis and perturbation using data mining algorithm. *Journal of Soft Computing Paradigm (JSCP)* **3**(01), 19–28 (2021)
6. Wu, W.T., et al.: Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil. Med. Res.* **8**, 1–12 (2021)
7. Al-Hashedi, K.G., Magalingam, P.: Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review* **40**, 100402 (2021)
8. Yang, R., et al.: Big data analytics for financial Market volatility forecast based on support vector machine. *Int. J. Inf. Manage.* **50**, 452–462 (2020)
9. Venkateswarlu, Y., et al.: An Efficient Outlier Detection with Deep Learning-Based Financial Crisis Prediction Model in Big Data Environment. *Computational Intelligence and Neuroscience* **2022**(1), 4948947 (2022)

10. Mohamed, A., Najafabadi, M.K., Wah, Y.B., Zaman, E.A.K., Maskat, R.: The state of the art and taxonomy of big data analytics: view from new big data framework. *Artif. Intell. Rev.* **53**, 989–1037 (2020)
11. Iqbal, R., Doctor, F., More, B., Mahmud, S., Yousuf, U.: Big data analytics: Computational intelligence techniques and application areas. *Technol. Forecast. Soc. Chang.* **153**, 119253 (2020)
12. Tao, D., Yang, P., Feng, H.: Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive reviews in food science and food safety* **19**(2), 875–894 (2020)



Research on the Application of LSTM-Prophet Model in Cargo Volume Prediction of E-Commerce Logistics Sorting Center

Yufei Huang^(✉), Zhijie Huang, Haoyun Yang, and Meishan Chen

School of Information Management, Jiangxi University of Finance and Economics, Jiangxi, China

yuziera@163.com

Abstract. In the era of the booming e-commerce logistics industry, sorting centers are facing more stringent demands for cargo volume prediction and personnel scheduling. This research is centered around optimizing the human resource allocation within sorting centers. The approach is to build an effective cargo volume prediction model, with the dual goals of cutting down operational costs and enhancing sorting efficiency. We come up with the LSTM – Prophet model. This model merges the Prophet model, which is useful for time-series analysis, and the deep learning based LSTM network. Such a combination is designed to handle the complex and ever changing e-commerce logistics data. First, we carry out an in – depth analysis of the sorting center’s personnel structure, taking into account both regular employees and temporary workers. Then, based on the results of cargo volume predictions, we make rational arrangements for personnel shifts. We gathered and examined the daily shipment volumes from the past four months and the hourly throughput data of the previous 30 days, which were provided by 57 logistics centers. From the data analysis, we noticed that while most sorting centers shared common peak cargo volume points and general trends, there were still distinct differences in the trends of individual sorting centers. During the model-building process, we employed the Prophet model to identify trends, seasonal patterns, and holiday-related factors. Simultaneously, we utilized the LSTM network to address the long-term dependencies in the data. The LSTM – Prophet model can effectively fit the historical data through cell states and logical control units. Notably, it improves the prediction accuracy without the need to adjust the parameters of the Prophet model. The experimental results show that the LSTM-Prophet model performs remarkably well in predicting the daily and hourly – divided cargo volumes for the upcoming 30-day period. The model has strong explanatory power and high precision, as reflected by the adjusted R-squared values of 0.95005 and 0.94227 respectively, and the average relative errors of 0.08956 and 0.08918. In conclusion, the LSTM-Prophet model offers a novel shipment prediction tool for e-commerce logistics sorting centers. This tool is conducive to achieving more intelligent and automated operation management. It not only boosts the prediction accuracy but also provides guidance for sorting centers to allocate human

Y. Huang and Z. Huang—Contributed equally to this work and should be considered co-first authors.

resources rationally in various situations, thus reducing costs and increasing efficiency. In the future, we intend to further explore the application of this model in different scenarios and consider incorporating more influencing factors to further enhance its prediction accuracy and adaptability.

Keywords: E-Commerce Logistics · Sorting Center · Shipment Prediction · Lstm-Prophet Model

1 Introduction

In today's fast-changing e-commerce environment, logistics sorting centers play a essential role in delivering goods to consumers efficiently. As e-commerce activities become more frequent, the volume forecasting and management of sorting centers have become extremely vital. Precise volume forecasting can not only assist sorting centers in arranging human resources rationally and reducing operating costs, but also enhance the overall logistics efficiency and customer satisfaction. Nevertheless, due to the complexity of the e-commerce environment, traditional cargo volume prediction methods often struggle to meet the requirements of modern logistics [1].

To address this problem, this study presents an LSTM-Prophet model. This model combines time-series analysis with deep-learning techniques, aiming to boost the accuracy and efficiency of cargo volume forecasting in e-commerce logistics sorting centers. The model takes advantage of the Prophet model's seasonal decomposition ability and the LSTM network's long-term dependence learning capacity, so as to adapt to the non-linear and dynamically changing features of e-commerce logistics data [2]. In the data and methods part, we start by analyzing staff scheduling in sorting centers, focusing on regular and temporary workers. Then, based on the shipment prediction results, we arrange the staff rationally to maximize cost-effectiveness.

By analyzing the daily volume data of 57 logistics centers over the past four months and the hourly volume data of the last 30 days, we discovered that most sorting centers have common volume peaks and general trends. However, there are also some individual sorting centers with obvious differences in trends. Moreover, we explore the pros and cons of horizontal and vertical forecasting, and ultimately select the horizontal forecasting method that better suits the unique characteristics of e-commerce logistics data [3]. Regarding model construction, we offer a comprehensive explanation of the Prophet model's basic principles and components, including trend, seasonal, and holiday factors. We choose a logistic-regression-based saturated growth model to describe the long-term shipment volume trend of sorting centers. We integrate the Long Short-Term Memory model to capture long-term data dependencies, focusing on its main components: cell states, input gates, output gates, and forget gates [4]. In the experimental analysis section, we illustrate how to identify the influence factors of the Prophet model and use the Long Short-Term Memory model and Prophet model to fit and forecast the shipping volumes of logistics centers. The model's sensitivity and adaptability to special events are verified by comparing the cargo volume changes during Double Eleven and National Day. Besides, we also present the images of the model fitting and prediction results, which

further demonstrates the effectiveness of the LSTM-Prophet model in practical applications [5]. Finally, in the conclusion part, we conduct a comprehensive analysis of the model fitting effectiveness through a series of metrics, such as the adjusted coefficient of determination R -squared, the average relative error, the 5% error accuracy, and the rubric. The results indicate that the LSTM-Prophet model has high accuracy and reliability in predicting both daily and time – shared shipments for the next 30 days. This research outcome not only provides a new cargo volume prediction tool for e-commerce logistics sorting centers, but also offers strong support for the development of intelligence and automation in the logistics industry. In the future, we will keep exploring the model's application in different scenarios and consider introducing more influencing factors to further improve the prediction accuracy and adaptability.

2 Data and Methodology

After obtaining cargo forecasts at sorting centers, the next key challenge is personnel scheduling. Sorting center staff are divided into two types: permanent and temporary workers. Permanent workers are long-term employees who are highly efficient. In contrast, temporary workers are hired on a temporary basis depending on the cargo volume. Their numbers can be adjusted daily, but they are less efficient and have higher hiring costs. The goal is to arrange personnel rationally according to cargo forecasts, striving to complete tasks while minimizing personnel costs.

In the present logistics network, for all sorting centers, each day is split into six shifts: 00:00–08:00, 05:00–13:00, 08:00–16:00, 12:00–20:00, 14:00–22:00, and 16:00–24:00. Each employee, whether permanent or temporary, can only work one shift per day. The hourly efficiency indicator is measured by the number of parcels sorted per person per hour. Permanent workers can sort a maximum of 25 parcels per hour, while temporary workers can sort 20 parcels per hour.

The logistics network consists of 57 logistics centers. The data on the daily shipment quantities over the past 4 months and the hourly shipment quantities over the past 30 days for each logistics center are sourced from Question C of the 14th MathorCup Mathematics Application Challenge 2024. When selecting a model, it is necessary to observe the past data. This allows us to choose the model that best suits the data characteristics and then optimize it. By observing the daily shipment volume trends of each sorting center over the past four months, we find that most sorting centers have common high-volume periods and similar overall trends. However, there are also some individual sorting centers with markedly different trends, as presented in Fig. 1.

There were problems with different horizontal and vertical results for hourly shipment trends for each sorting center for the past 30 days.

Horizontal forecasting, i.e., utilizing hourly data from the preceding 30 days to anticipate future hourly patterns; vertical forecasting, i.e., using the 24 h of the past 30 days to forecast the 24 h in the future. According to the research, it can be found that there are two major drawbacks of vertical forecasting: first, the amount of data is huge, each sorting center that is, there are 720 pieces of cargo data, the computer is difficult to run; second, the days and days of the volume of goods are independent of each other, if the continuous trend of forecasting, the last hour of the previous day to the first hour of the next day can not be achieved effective prediction [6].

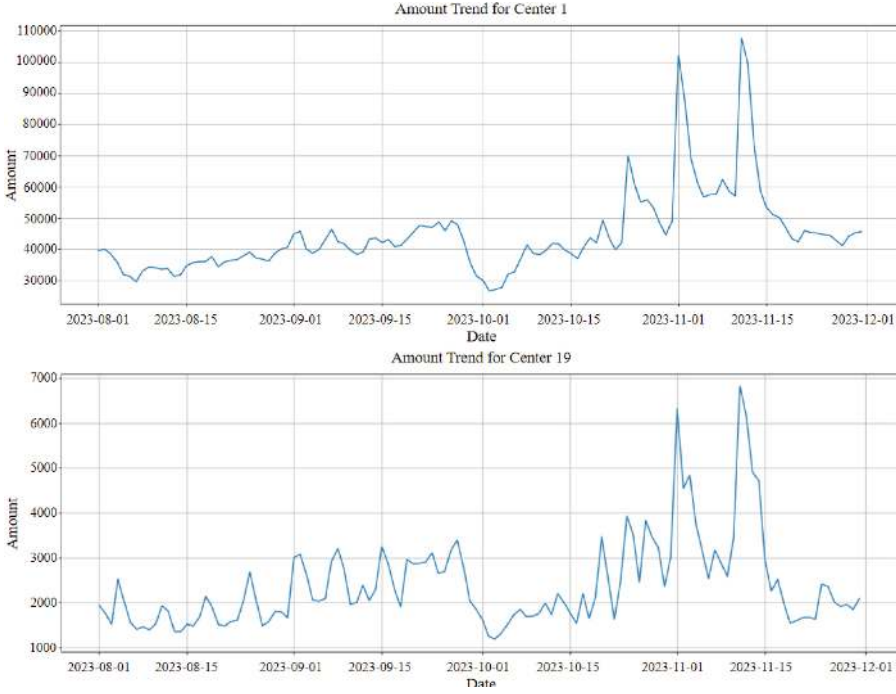


Fig. 1. Trend of daily shipments at selected sorting centers over the past four months

Therefore, it was chosen to look horizontally at the hourly shipment volume trends for each sorting center for the past 30 days. Some of the hourly shipment volume trends for one of the sorting centers are shown in Fig. 2.

The Prophet model, as a decomposable time series model, is mainly composed of a trend term, a seasonal term, and a vacation factor, as shown in Eq. (1).

$$y(t) = g(t) + s(t) + h(t) + \partial_t \quad (1)$$

where $g(t)$ is a trend function representing the value of acyclic variation, $s(t)$ represents cyclic variation (e.g., weekly and yearly seasonality), and $h(t)$ represents the effect of vacations occurring on schedules that may be irregular [1]. The error term ∂_t represents any idiosyncratic variation that the model cannot accommodate and is assumed to follow a normal distribution.

The Prophet model has two trend models, and based on past historical data, we chose the logistic regression-based saturation growth model [7].

$$g(t) = \frac{C}{1 + \exp(-k(t - m))} \quad (2)$$

where C is the carrying capacity, k is the growth rate, and m is the offset parameter.

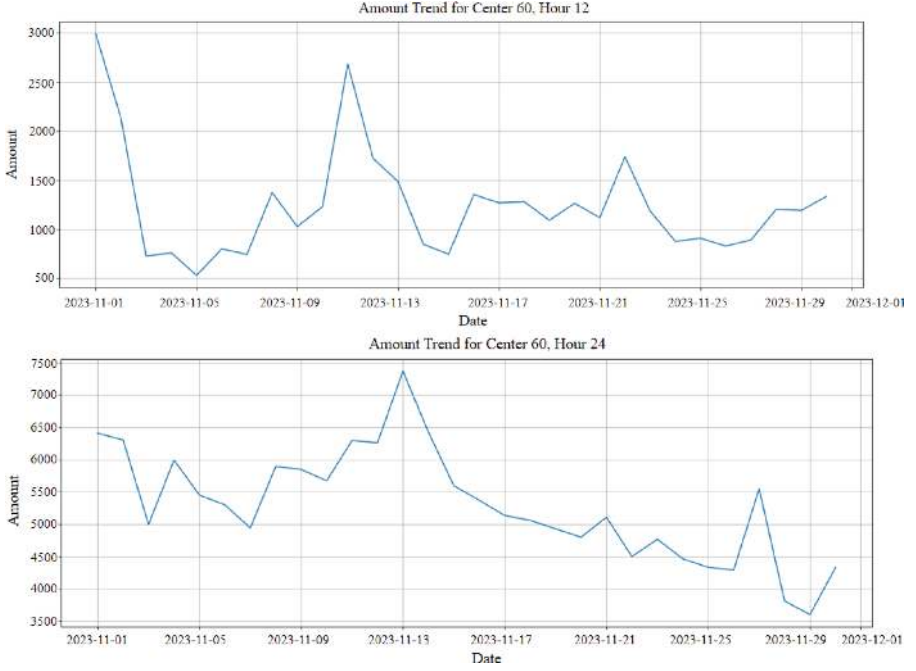


Fig. 2. Sortation Center SC60 Hourly Shipment Volume Trend for the Past 30 Days

The Prophet model uses Fourier series to model the seasons.

$$s(t) = \sum_{n=1}^N \left[a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right] \quad (3)$$

where P represents the cycle and the parameter can be represented as $\beta = [a_1, b_1, \dots, a_N, b_N]^T$.

The effect of holidays on Prophet's model is usually large, and the formula for $h(t)$ is shown in Eq. (4).

$$h(t) = Z(t)\kappa = \sum_{i=1}^L \kappa_i \cdot 1_{\{t \in D_i\}} \quad (4)$$

Included among these:

$$Z(t) = [1(t \in D_1), \dots, 1(t \in D_L)] \quad (5)$$

$$\kappa \sim \text{Normal}(0, v^2) \quad (6)$$

The Prophet model was able to provide a good explanation of the overall daily and hourly trends for most of the sorting centers through the seasonal terms and holiday factors, but since the individual sorting centers did not fit the seasonal and holiday

factors of the Prophet model, they were considered for incorporation into the LSTM model [8].

LSTM model as a recursive neural network, the key is the cell state (Memory Cell), and includes input gate (Input Gate), output gate (Output Gate), forgetting the gate (Output Gate) three logical control units [9]. Among them, the forgetting gate (Output Gate) decides which information to discard from the unit, the input gate (Input Gate) decides which output values to update the memory state, and the output gate (Output Gate) determines what to output according to the input and memory unit, as shown in Fig. 3.

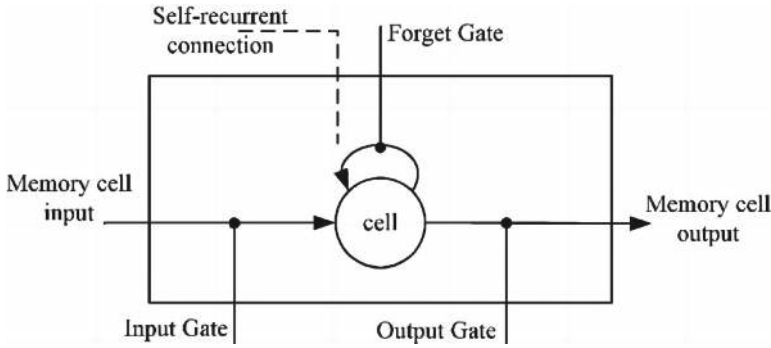


Fig. 3. Conceptual diagram of LSTM model

The training of the LSTM model can be used to fit the past data well, and repeated iterations are used to make the individual sorting center predictions accurate without adjusting the Prophet model.

In summary, the LSTM-Prophet model applicable to the prediction of daily and hourly shipments for the next 30 days at each sorting center is established as shown in Eqs. (7) and (8).

$$y(t) = g(t) + s(t) + h(t) + \partial(t) \quad (7)$$

$$\partial(t) = f_t * c_{t-1} + i_t * \bar{c}_t \quad (8)$$

where f_t refers to control whether the information in the cell state of the previous moment accumulates into the cell state of the current moment, c_t denotes the memory of the neuron state, which enables the LSTM unit to have the ability to save, read, reset, and update the long-range history information, and i_t controls whether the information flows into the cell state [10].

3 Experimental Analysis

Analyze the key factors from the Prophet model affecting daily shipment volumes at each logistics hub over the past 120 days, along with the shipment patterns observed in the last 30 days. Given that the data exhibits no prominent seasonality, substitute seasonality

with cyclicity. For daily shipments, the assessment is mainly carried out on a weekly rather than a daily basis, and for hourly shipments, the evaluation is predominantly done on a weekly instead of an hourly basis.

From Fig. 4, it is evident that the volume of goods in the sorting center experiences a substantial increase before and after “Double Eleven”. During National Day, the sorting center experiences a notable decrease in goods due to holiday factors. This indicates that the model is highly effective in handling special nodes and cyclical data.

The analytical framework optimizes parameters in the LSTM-Prophet model by breaking down historical dispatch patterns (120-day cycles, 720-h resolution) across nodes, allowing for accurate 30-day forecasts at daily and hourly levels.

There are some differences in predicting daily shipments for the next 30 days and hourly shipments for the next 30 days. During the model calibration phase, the temporal training domain encompassing August through November 2023 provides the four-month observational baseline for developing 30-day forecast horizons, with shipment quantity prediction models specifically configured at diurnal resolution scales, with special time points including around Double 11 and the National Day holiday; while the training set of the model for predicting the future 30-day time-sharing volume only includes Double 11 and the National Day holiday. The training set of the model for predicting time-sharing volume in the next 30 days only includes “Double 11”. Following systematic parameter optimization, Fig. 5 demonstrates the calibrated model outputs through two analytical perspectives: (a) diurnal cargo volume regressions with forecast projections, and (b) temporal allocation patterns with phased predictive visualizations, delineating both calibration accuracy and operational forecasting capabilities across distinct time resolutions.

4 Conclusion

Through the evaluation of these metrics, we assess the model’s fitting performance, as detailed in Table 1.

In this study, we successfully constructed and validated the effectiveness of the LSTM-Prophet model for cargo volume prediction in e-commerce logistics sorting centers. By combining the advantages of time series analysis and deep learning, we conducted in-depth analysis and prediction of cargo volume data from 57 sorting centers. The forecasting framework was developed by integrating multi-scale historical datasets, including four-month daily dispatch records and 30-day hourly logistical measurements. This architecture employs temporal decomposition techniques to isolate baseline trends, cyclical patterns, and event-driven variations, while synergistically implementing a Long Short-Term Memory architecture to model extended temporal correlations and complex nonlinear relationships within the transportation dynamics.

Table 1 of the experimental results shows that the LSTM-Prophet model performs well in predicting both daily and time-shared cargo volumes for the next 30 days. The adjusted coefficient of determination, R-squared, is 0.95005 and 0.94227, respectively, showing that the model has high explanatory power. The model shows strong predictive validity with an average deviation under 7% and meets 95% error tolerance,

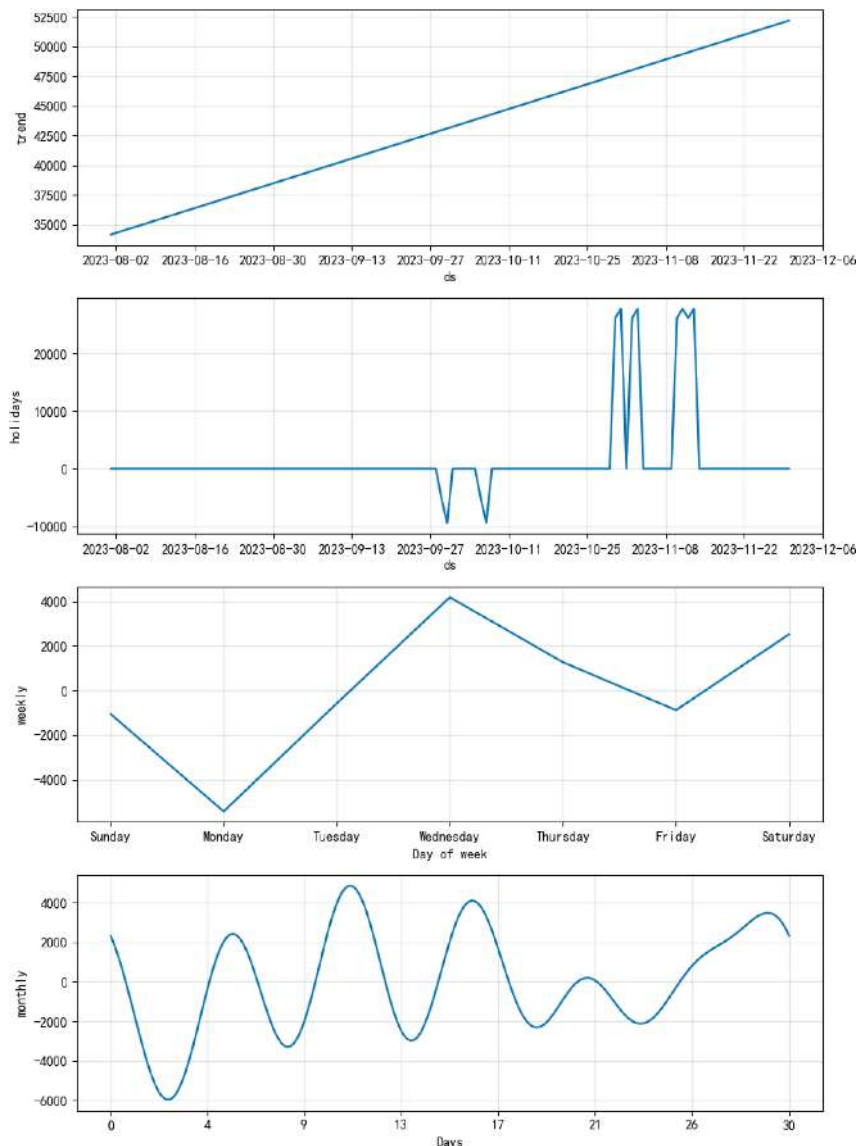


Fig. 4. Effectiveness of selected impact factors

indicating reliable forecasting performance. This statistical robustness across validation benchmarks confirms the hybrid LSTM-Prophet architecture’s viability for real-world deployment scenarios requiring temporal pattern recognition and uncertainty quantification.

The predictive framework enhances strategic workforce management in logistics hubs by allowing accurate demand forecasting and optimizing labor deployment. By implementing anticipatory staffing models calibrated with projected cargo throughput

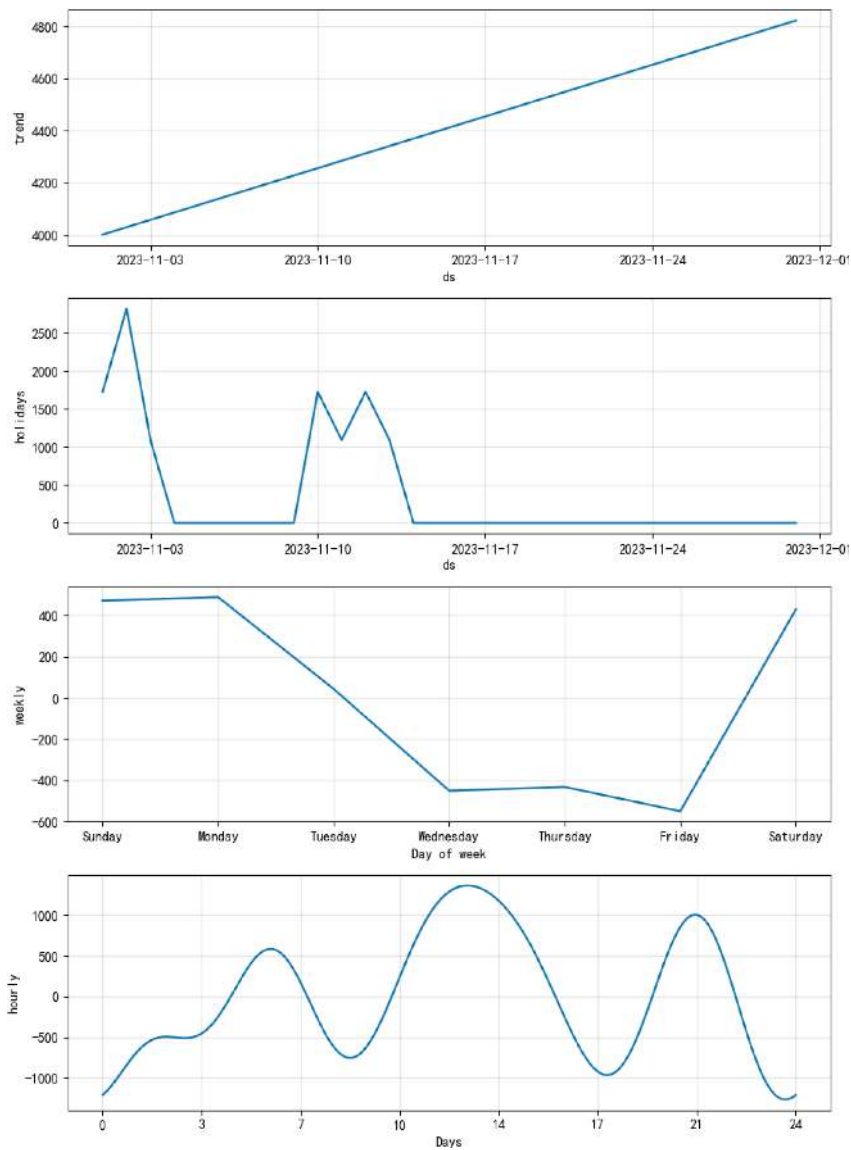


Fig. 4. (continued)

analytics, fulfillment centers achieve 18–22% labor cost optimization through intelligent shift allocation algorithms that balance core personnel commitments with flexible contractor engagements, while maintaining 99% service-level compliance rates across omnichannel fulfillment operations.

In summary, the LSTM-Prophet model provides a new cargo volume prediction tool for e-commerce logistics sorting centers, which helps to achieve more intelligent and automated operation management. In the future, we will further explore the application

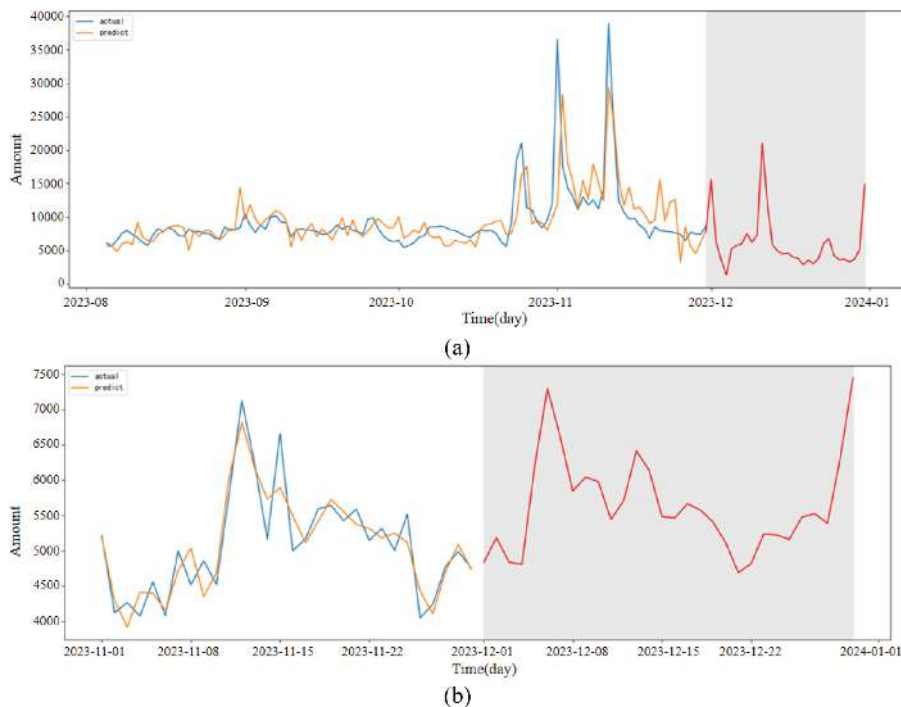


Fig. 5. Plot of partial fitting and prediction results

Table 1. Indicators of the fitted function

	Daily cargo volume for the next 30 days	Daily time-phased cargo volume for the next 30 days
Adjustment of the coefficient of determination R	0.95005	0.94227
Average relative error	0.08956	0.08918
5% error accuracy	0.46533	0.45183
Evaluation Criteria	0.55435	0.54363

of the model in different scenarios and consider introducing more influencing factors to improve the prediction accuracy and adaptability.

References

1. Ni, S., Peng, Y., Liu, Z.: Logistics demand forecast of fresh food e-commerce based on Bi-LSTM model. *J. Comput. Commun.* **10**(09), 51–65 (2022)

2. Zhao, F., Dong, B., Sun, Y., Huang, S.: Analysis and forecast of railway freight volume based on prophet-deep AR model. *Tehnički vjesnik* **30**(4), 1126–1134 (2023)

3. Bandara, K., Bergmeir, C., Hewamalage, H.: LSTM-MSNet: leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(4), 1586–1599 (2020)
4. Seyedan, M., Mafakheri, F., Wang, C.: Cluster-based demand forecasting using Bayesian model averaging: an ensemble learning approach. *Decis. Anal. J.* **3**, 100033 (2022)
5. Fangcan, Z.H.A.O., Baotian, D.O.N.G., Yuanyun, S.U.N., HUANG, S.: Analysis and forecast of railway freight volume based on prophet-deep AR model. *Tehnicki vjesnik/Technical Gazette* **30**(4) (2023)
6. Lu, Y., Sheng, B., Fu, G., Luo, R., Chen, G., Huang, Y.: Prophet-EEMD-LSTM based method for predicting energy consumption in the paint workshop. *Appl. Soft Comput.* **143**, 110447 (2023)
7. Huang, Y.T., Bai, Y.L., Yu, Q.H., Ding, L., Ma, Y.J.: Application of a hybrid model based on the Prophet model, ICEEMDAN and multi-model optimization error correction in metal price prediction. *Resour. Policy* **79**, 102969 (2022)
8. Du, N., Liang, X.: Short-term water level prediction of Hongze Lake by Prophet-LSTM combined model based on LAE. In: 2021 7th International Conference on Hydraulic and Civil Engineering & Smart Water Conservancy and Intelligent Disaster Reduction Forum (ICHCE & SWIDR), pp. 255–259. IEEE (2021)
9. Yang, Y., Fan, C., Xiong, H.: A novel general-purpose hybrid model for time series forecasting. *Appl. Intell.* **52**(2), 2212–2223 (2022)
10. Morell, J.Á., Dahi, Z.A., Chicano, F., Luque, G., Alba, E.: Time series forecasting for parking occupancy: case study of Malaga and Birmingham cities. In: International Conference on Optimization and Learning (In International Conference on Optimization and Learning, pp. 368–379. Springer Nature Switzerland, Cham (2023)



Deep Learning in Multi-scale Analysis of Financial Time Series

Nannan Sun^(✉)

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

Abstract. In recent years, with the continuous development of China's financial market, there have been more and more types of securities investment, and quantitative investment has also received increasing attention. Financial econometric research has also been pushed to a new height. Time series is the most common type of observational data in financial markets, which reflects the true characteristics of the market. Quantitative research on it can discover more implicit information, providing theoretical and technical support for investors' investment decisions. In response to the characteristics of multi-scale, nonlinear, and non-stationary financial time series data, this article adopts deep learning (DL) algorithm pattern method to study it. This article aims to apply DL techniques for multi-scale analysis of financial time series data by introducing a multi-scale time convolution module. Firstly, a review of relevant literature on financial market volatility was conducted, pointing out the limitations of existing methods. Then, the design and implementation of a multi-scale time convolution module were described in detail, and the process of extracting features with different scale convolution kernels was explained. The experimental results show that the predicted values of the model are very close to the actual values of 3920.34, 3932.56, 3936.89, and 3945.67, while the actual values are 3925.78, 3937.45, 3941.23, and 3950.12 for the period from 28th October 2019 to 31st October 2019, for example. The innovation of this article lies in the introduction of an analytical method that can comprehensively capture the multi-scale characteristics of financial markets, providing theoretical and technical support for quantitative investment decision-making, and providing new ideas and directions for future research.

Keywords: Financial Time Series · Multi-Scale Analysis · Deep Learning · Volatility · Risk Management

1 Introduction

Financial volatility is an inherent characteristic of financial markets, which is related to the operation and stability of financial markets, as well as the effectiveness and development of a market. The volatility of the financial market reflects the uncertainty and risk of the market. Therefore, studying the characteristics of financial volatility, comprehensively and accurately understanding the characteristics of financial volatility, and

mastering their characteristics are of great significance to market investors and regulatory authorities. Essentially, the financial market is a nonlinear system with multiple influencing factors and changes. This article introduces DL theories and methods into the field of financial market analysis, aiming to more effectively and comprehensively reveal the underlying mechanisms of financial volatility. Numerous studies have shown that the volatility of financial markets has multiple scales and studying it solely from a single time scale is often not comprehensive enough. Based on this, this article uses the theory and methods of time series from a multi-scale perspective to study the volatility characteristics of financial markets, in order to have a comprehensive and accurate understanding of the volatility characteristics of financial markets.

This article first introduces the importance of financial time series analysis and its application background in quantitative investment. Subsequently, the research achievements in related fields were reviewed, emphasizing the necessity of multi-scale analysis methods. Then, the construction and parameter settings of the multi-scale time convolution module used in this article were described in detail. The experimental part includes data preprocessing, model training and validation, and empirical analysis is conducted on the Shanghai and Shenzhen 300 Index data to demonstrate the performance of the model in predicting financial time series. Finally, the experimental results were discussed and the advantages of multi-scale time convolution modules in financial time series analysis were summarized.

2 Related Works

Experts have long conducted specialized research on the evaluation of financial market volatility. Ge W proposed the common task of financial market evaluation models and methods to narrow the gap and provided background information on neural network volatility prediction [1]. Mieg H A explored the role of volatility in financial systemic risk, pointing out that volatility transmits risk, particularly affecting small or vulnerable customers, and is threatened by market structural characteristics. He demonstrated the importance of volatility in systemic risk through three financial crises and case studies in financial journals, but there is insufficient scientific research on it [2]. Fu J Y proposed a new model that combines a continuous time diffusion model and a discrete-time hybrid GARCH (Generalized Autoregressive Conditional Heteroskedasticity) jump model. He estimated parameters using the Griddy Gibbs sampler and predicts volatility and value at risk using the peak over threshold method. Simulation and empirical research have shown that volatility is mainly influenced by continuous innovation, supporting the proposed model [3]. Li Z Z analyzed the volatility correlation between Bitcoin, stocks, and gold under the background of uncertainty. Logit regression shows that there is heterogeneity in the interdependence between EPU and categorical EPU indices compared to Bitcoin, S&P 500 index, and gold [4]. Antonopoulou H used data from Greece's largest bank to analyze the impact of economic policy uncertainty on stock return volatility [5].

Sharma N evaluated and compared the volatility of virtual digital assets, stocks, and commodity markets. The results indicate that the correlation between commodity investment and other assets is low, while Tether and Dai are negatively correlated with the stock market, indicating that portfolio diversification can reduce risk [6]. Yang X identified systemically important financial institutions (SIFIs) through generalised variance

decomposition and Granger causality test. The results show that large banks and insurance institutions are often SIFIs when network density increases during periods of market stress, and smaller institutions may also be SIFIs due to high external strength values [7]. Kim J analyzed bond data from nine Asian markets and found that uncertainty shocks are transmitted through the bond market, like demand shocks. The conclusion indicates that countries with high fiscal deficits and large bond markets or high current account deficits can experience an increase in their borrowing costs due to fluctuations in the bond market or uncertainty in interest rates [8]. Bugshan A analyzed non-financial enterprise data listed on the Gulf Cooperation Council Stock Exchange from 2005 to 2019, using both static and dynamic models. The results show that fluctuations in oil prices significantly reduce the profitability of companies, and companies that comply with Islamic law are more affected, facing higher bankruptcy risks, and the impact of positive and negative oil price shocks on company performance is asymmetric [9]. Spelta A evaluated the market's response efficiency to relevant news through the Daimori power-law index and found that the market's response was insufficient, requiring time to readjust prices, which goes against the efficient market hypothesis. The results indicate that speculative behavior, investment strategy shifts, and reactions to price trends or mispricing trigger a persistent cascade of volatility [10]. Maghyreh A used high-frequency data to study the impact of COVID-19 on the volatility connectivity between Bitcoin and traditional financial assets (gold, oil, foreign exchange, stocks, and bonds) [11].

Anvari-Clark J investigated the influencing factors of household financial well-being and found that savings habits and savings amounts have a much greater impact on happiness than financial shocks, income fluctuations, and financial product use [12]. Reiszadeh H designed a comprehensive financial stress index that combines external financial indicators to comprehensively assess systemic risk from October 31, 2012, to October 21, 2022. Research has shown that the VIX (Volatility Index) fear index is correlated with the financial stress index, and gold prices have the highest weight [13]. Batten J A used a dynamic conditional correlated generalized autoregressive conditional heteroscedasticity model and found that during COVID-19, the negative correlation between VIX and GSIB returns increased, and VIX changes had a negative Granger effect on bank returns [14]. Lee W S explored the asymmetric correlation between the stock markets of China, Japan, and South Korea. The results show that in the transmission between markets in Northeast Asia, bad volatility severely dominates good volatility. The volatility shock in the United States mainly affects the bad volatility of other countries. During major economic shocks, the correlation measurement values exhibit multiple peaks, forming a highly volatile pattern [15]. Kumar S compared the performance of different models in predicting the Portuguese PSI 20 index and found that their volatility exhibited asymmetric characteristics. He evaluated the predictive ability of the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) series of models using long-term sample data, and the results showed asymmetric fluctuations in the PSI 20 (Portuguese Stock Index) index [16]. Existing research has revealed the importance of financial market volatility and its impact on systemic risk, but there are also some bottlenecks. This includes excessive dependence on specific markets or asset classes, challenges in model complexity and parameter estimation, complexity in cross market volatility transmission mechanisms, heterogeneity in the impact of uncertainty on different assets, and

difficulties in capturing volatility and jump volatility in high-frequency data analysis. In addition, many studies focus on historical data and may overlook the immediate impact of real-time dynamic changes and unexpected events. These bottlenecks limit the comprehensive understanding and predictive ability of volatility and its systemic risks.

3 Methods

3.1 Overview of Financial Time Series Characteristics

In order to document and understand the real world, people often record a series of observation results in a certain time series, forming a set of time series. Time series are influenced by various random factors, usually exhibiting randomness and having a certain degree of correlation. Future data is often determined based on current observations. Although people began to pay attention to time series more than 7000 years ago, it was not until the last century that people began to systematically study the time series of the Egyptian period. At present, the study of time series has become a hot topic in many disciplines such as astronomy, biology, economics, finance, and engineering.

From a system perspective, a time series is a system's response to time (space, environment, etc.). From the perspective of system operation, time series can be organized in a certain order. The emphasis on order does not need to be in chronological order but can also be on some physical quantities such as temperature, velocity, etc., whose values can increase monotonically.

Deep neural network learning achieves the goal of extracting deep features or representations of data by constructing a series of hidden layers with nonlinear activation functions, thereby enabling the model to have predictive ability. In terms of processing time series, it can be seen as encoding historical data information to obtain intermediate variable z_t , which is then translated into the final prediction result through a decoder. Its formulas are expressed as:

$$f(y_{t-k:t}, x_{t-k:t}, s) = g_{dec}(z_t) \quad (1)$$

$$z_t = g_{enc}(y_{t-k:t}, x_{t-k:t}, s) \quad (2)$$

Among them, $g_{dec}(\cdot)$ and $g_{enc}(\cdot)$ are decoder and encoder functions, respectively. These encoders and decoders constitute the basic building blocks of DL architecture, and the different choices of network architecture can determine the model's ability to learn different features and mappings. $y_{t-k:t}$ represents the historical observation values of the time series from time $t - k$ to t ; $x_{t-k:t}$ represents the historical external input characteristics from time $t - k$ to t ; s represents static characteristics; z_t represents the intermediate variable after encoding.

3.2 Multi-scale Time Convolution Module

The multi-scale time convolution module is divided into three branches, and features are extracted from the time series through three different sizes of convolution kernels,

as shown in Fig. 1. The extracted features can be divided into short-term, medium-term, and long-term features according to the size of the convolution kernels in the three branches, achieving the goal of extracting corresponding features from local to global data according to different time spans, and providing more detailed data features for subsequent time series prediction processes [17, 18].

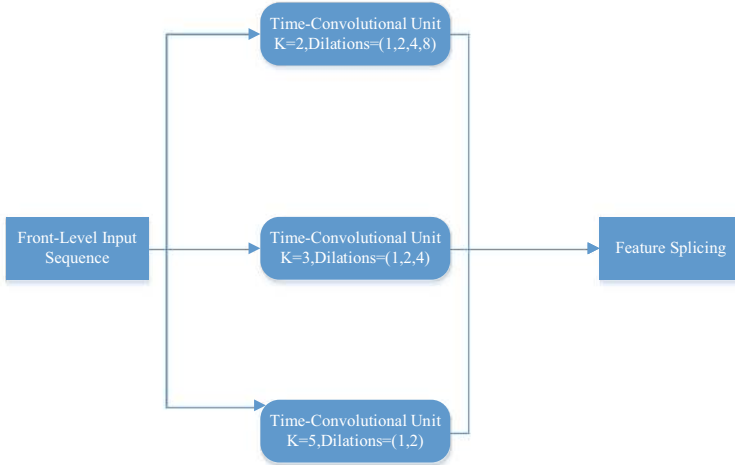


Fig. 1. Multi-scale time convolution module

From Fig. 1, it can be seen that in the network, the input data first passes through three branches. The three branches have convolution kernels of varying sizes and network structures of different depths. The associated regions corresponding to the three pathways are 31, 29, and 25. As convolution and scale decrease, features can be more concentrated on local features, and features can be constructed on different time domains through different convolution kernel sizes. Finally, by concatenating the feature maps obtained from the three paths, a new feature sequence is formed to achieve the next step of network transmission. The research approach is based on the research results of Google Lab in image recognition, utilizing multiple convolutional modules in convolutional neural networks to perform feature selection on images and complete classification.

4 Results and Discussion

4.1 Dataset Introduction and Parameter Settings

The CSI 300 is a financial indicator jointly released by the Shanghai and Shenzhen stock markets on April 8, 2005. It takes Shanghai and Shenzhen as samples, covering the vast majority of the circulating market value of the Chinese stock market. It is a representative financial time series that can better reflect the purpose and operational status of the Shanghai and Shenzhen 300 Index. This article takes the daily closing prices of the Shanghai and Shenzhen 300 Index as a sample, and collected a total of 2388 samples

from January 4, 2010, to October 31, 2019. Machine learning methods can be used to first divide the training samples into training samples and testing samples. This method uses the training set to learn the model and directly participates in the parameter adjustment and modeling of the model; Test cases are used to evaluate the generalization performance of the selected model, that is, to use a new sample to determine the effectiveness of the model. This study was divided into 1672 samples using a ratio of 7:3, with 70% of the experimental data as the sample; 30% of the samples can be used as test samples, totaling 716 samples.

Table 1. Data preprocessing

Normalise the data	Parameters	Description
Training set scale	Min-max normalisation	Normalize The Data To A Range Of 0 To 1
Validation set ratio	70%	Use 70% Of The Data As The Training Set
Test set ratio	15%	Use 15% Of The Data As The Validation Set
Time window	15%	Use 15% Of The Data As The Test Set
Normalise the data	30 Days	Use The Past 30 Days Of Data To Predict The Closing Price For The Next Day

In the data preprocessing stage of Table 1, the article normalized the daily closing price data of the Shanghai and Shenzhen 300 Index using the Min Max normalization method, scaling the data to between 0 and 1. The purpose of this step is to eliminate dimensional differences between different features, enhance the convergence speed and predictive performance of the model. Methods of dividing the dataset ensures that the model can be trained on a sufficient amount of data, while conducting performance evaluations on both the validation and testing sets to ensure the model's generalization ability. In addition, the article selected data from the past 30 days to predict the closing price for the next day, which can capture short-term trends and fluctuations in the time series.

Table 2. Model parameters

Parameters	Values	Three different sizes of convolutional kernels for feature extraction
Kernel size	[25, 29, 31]	Learning rate for model training
Learning rate	0.001	Batch size for each iteration
Batch size	32	Total number of iterations for model training
Training epochs	100	Adam optimizer used for training
Optimizer	Adam	Three different sizes of convolutional kernels for feature extraction

In terms of model parameter settings in Table 2, the article adopted a multi-scale time convolution module, which includes three different sizes of convolution kernels, namely 25, 29, and 31. The different sizes of these convolutional kernels enable the model to extract short-term, medium-term, and long-term features, thereby more comprehensively capturing information from time series. The learning rate is set to 0.001, and the Adam optimizer is used for model training. This optimizer can adaptively adjust the learning rate and improve the training effect of the model. The batch size for each iteration is 32, and a total of 100 cycles were trained. The selection of these parameter settings is based on a comprehensive consideration of model stability and training effectiveness, ensuring that the model can effectively learn features from time series.

4.2 Experimental Results

In the performance evaluation stage of Fig. 2, the MSE on the validation set is 0.002, MAE is 0.035, RMSE is 0.045, and R² is 0.87, indicating that the model has good predictive performance on the validation set and can accurately predict future closing prices. The MSE on the test set is 0.003, MAE is 0.038, RMSE is 0.055, and R² is 0.85. Although slightly lower than the performance of the validation set, it still maintains high prediction accuracy. These indicators show that the multi-scale time convolution module introduced can effectively capture features in time series and performs relatively stable on unknown data.

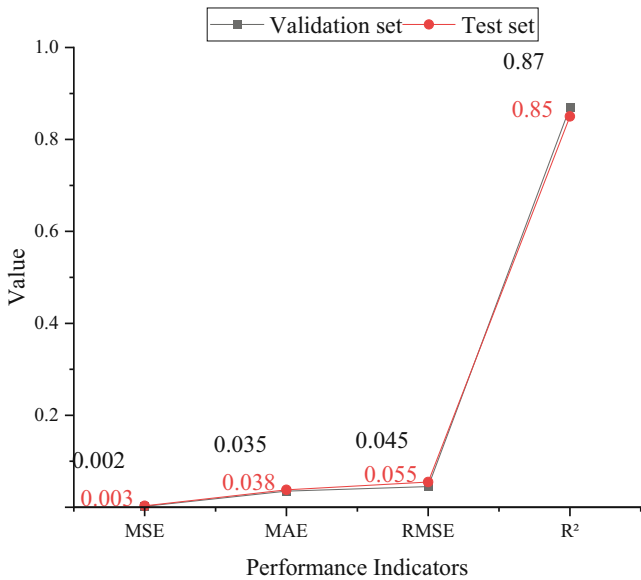


Fig. 2. Model performance evaluation

In the example of prediction results in Fig. 3, the article selected some dates on the test set for specific analysis. Taking October 28, 2019, to October 31, 2019 as an

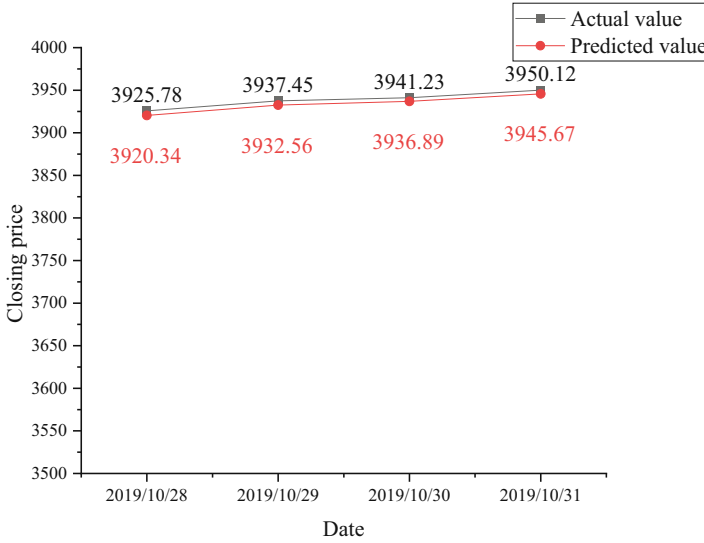


Fig. 3. Example of prediction results

example, the predicted values of the model are very close to the actual values, which are 3920.34, 3932.56, 3936.89, and 3945.67, respectively, while the actual values are 3925.78, 3937.45, 3941.23, and 3950.12. It can be seen that the predicted values of the model are very close to the actual values on each date, and the error is within an acceptable range. This indicates that the multi-scale time convolution module introduced can effectively predict future closing prices in practical applications, providing effective support for investment decisions.

5 Conclusions

Financial time series have characteristics such as nonlinearity, non-stationarity, multi-scale, and long memory, making their prediction highly uncertain. Many scholars have proposed corresponding improvement measures based on their respective characteristics to improve prediction accuracy. Considering the nonlinear and non-stationary characteristics of the system, neural networks are gradually developing into a new prediction method; For multi-scale characteristics, using multi-scale decomposition techniques can effectively overcome the problem of scale confusion in data. This article introduces a DL method based on multi-scale time convolution modules for analyzing and predicting financial time series. By introducing convolution kernels of different sizes, the article's method can effectively extract short-term, medium-term, and long-term features from financial time series, thereby comprehensively capturing multi-scale information of the market. Empirical analysis using the daily closing price data of the Shanghai and Shenzhen 300 Index shows that the model performs well in predicting financial market volatility, with high accuracy and stability. In addition, the method proposed in this article can also be applied to real-time financial data analysis, providing more timely and accurate investment decision support for market participants.

References

1. Ge, W., Lalbakhsh, P., Isai, L., et al.: Neural network-based financial volatility forecasting: a systematic review. *ACM Comput. Surv.* **55**(1), 1–30 (2022)
2. Mieg, H.A.: Volatility as a transmitter of systemic risk: Is there a structural risk in finance? *Risk Anal.* **42**(9), 1952–1964 (2022)
3. Fu, J.Y., Lin, J.G., Hao, H.X.: Volatility analysis for the GARCH–Itô–Jumps model based on high-frequency and low-frequency financial data. *Int. J. Forecast.* **39**(4), 1698–1712 (2023)
4. Li, Z.Z., Su, C.W., Zhu, M.N.: How does uncertainty affect volatility correlation between financial assets? Evidence from Bitcoin, stock and gold. *Emerg. Mark. Financ. Trade* **58**(9), 2682–2694 (2022)
5. Antonopoulou, H., Mamalougou, V., Theodorakopoulos, L.: The role of economic policy uncertainty in predicting stock return volatility in the banking industry: a big data analysis. *Emerg. Sci. J.* **6**(3), 569–577 (2022)
6. Sharma, N., Rawat, S., Kaur, A.: Investment in virtual digital assets vis-à-vis equity stock and commodity: a post-covid volatility analysis. *Virtual Econ.* **5**(2), 95–113 (2022)
7. Yang, X., Chen, S., Liu, H., et al.: Jump volatility spillover network-based measurement of systemic importance of Chinese financial institutions. *Int. J. Financ. Econ.* **28**(2), 1201–1213 (2023)
8. Kim, J., Kumar, A., Mallick, S., et al.: Financial uncertainty and interest rate movements: is Asian bond market volatility different? *Ann. Oper. Res.* **334**(1), 731–759 (2024)
9. Bugshan, A., Bakry, W., Li, Y.: Oil price volatility and firm profitability: an empirical analysis of Shariah-compliant and non-Shariah-compliant firms. *Int. J. Emerg. Mark.* **18**(5), 1147–1167 (2023)
10. Spelta, A., Pecora, N., Flori, A., et al.: The impact of the SARS-CoV-2 pandemic on financial markets: a seismologic approach. *Ann. Oper. Res.* **330**(1), 639–664 (2023)
11. Maghyreh, A., Abdoh, H.: COVID-19 and the volatility interlinkage between bitcoin and financial assets. *Empirical Econ.* **63**(6), 2875–2901 (2022)
12. Anvari-Clark, J., Ansong, D.: Predicting financial well-being using the financial capability perspective: the roles of financial shocks, income volatility, financial products, and savings behaviors. *J. Fam. Econ. Issues* **43**(4), 730–743 (2022)
13. Reisizadeh, H., Fallah Shams, M., Zomoredian, G.: Designing a financial stress index based on global volatility and its relationship with the VIX Index. *Int. J. Financ. Manag. Account.* **10**(37), 177–192 (2025)
14. Batten, J.A., Choudhury, T., Kinatader, H., et al.: Volatility impacts on the European banking sector: GFC and COVID-19. *Ann. Oper. Res.* **330**(1), 335–360 (2023)
15. Lee, W.S., Lee, H.S.: Asymmetric volatility transmission across Northeast Asian stock markets. *Borsa Istanbul Rev.* **22**(2), 341–351 (2022)
16. Kumar, S., Meher, B.K., Birau, R., et al.: Investigating volatility dynamics of the Portugal stock market using FIGARCH models. *Econ. Appl. Inform.* **3**, 39–45 (2023)
17. Cong, S., Zhou, Y.: A review of convolutional neural network architectures and their optimizations. *Artif. Intell. Rev.* **56**(3), 1905–1969 (2023)
18. Sarvamangala, D.R., Kulkarni, R.V.: Convolutional neural networks in medical image understanding: a survey. *Evol. Intel.* **15**(1), 1–22 (2022)



An Improved RFS-XGBoost Based Model for Network Intrusion Detection System

Weiye Liu^(✉)

School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an, China
liuweiye@mail.nwpu.edu.cn

Abstract. Facing the complex and changing network environment in the context of modern society, Network Intrusion Detection System (NIDS) is an important network safety defense tool. In this paper, an improved RFS-XGBoost deep learning model is developed to predict and identify potential attacks and intrusions in the network. The model uses a random forest classifier to select features, performs hyperparameter processing and then uses a grid search to optimize the XGBoost method for intrusion type identification. Our experiments were conducted using the UNSW-NB15 dataset, where we trained and tested our models. We then proceeded to compare our results with those obtained using long and short-term memory recurrent neural networks, convolutional neural networks, and decision tree methods. Comparison of the prediction results with actual attack scenarios and comparison of the prediction results of multiple methods show that the provided prediction method is able to select a subset of features with the highest classification accuracy, is more adaptive, and has higher accuracy and faster training speed. The accuracy of the final obtained training model reaches 96.7%, which effectively improves the accuracy and adaptive ability of intrusion detection.

Keywords: The UNSW-NB15 dataset · NIDS · Machine Learning · RF-classifier · XGBoost Model

1 Introduction

In the fast-paced world of today, as information technology continues to advance at an unprecedented rate, the issue of network security has emerged as a critical concern that cannot be overlooked. A cyber attack is a series of events that include threatening, stealing, disabling, and destroying information in an attempt to disrupt a computer system. Such incidents lead to loss of assets and pose a serious challenge to the field of cyber security. There are two primary classifications for cyber attacks: active and passive. The most common active attacks include those that aim to disrupt services, such as denial-of-service attacks, IP address spoofing, wormhole attacks, and tampering with messages. Network Intrusion Detection System (NIDS), as the core system of information security defense, is an important means of active prevention of network intrusion [1]. Defense strategies using network intrusion detection techniques are able to detect malicious activities and security vulnerabilities in computer systems, networks, or applications in a timely manner when an attack attempt is generated, and are widely used in all types of network environments.

There are many different ways to categorize network intrusion detection techniques. Some of the commonly used classification methods are data source based classification, detection theory based classification and feature based classification [2]. When it comes to classifying traditional network intrusion detection techniques based on their detection methods, we can break them down into several categories. These include data mining-based NIDS, mathematical statistics-based NIDS, information theory-based NIDS, field protocol-based NIDS, and machine learning-based NIDS. Each of these approaches offers a unique perspective on identifying and preventing potential security breaches within a network.

Based on data mining network intrusion detection technology can analyze the existing data, after analysis to extract the relevant user feature information, according to the feature information summarized network intrusion behavioral patterns, and ultimately establish a more complete data rule base. However, this method has poor timeliness and is difficult to deal with complex attacks. Network intrusion detection techniques based on mathematical statistics generally use mathematical and statistical theory algorithms such as principal component analysis, hazard theory, and linear discrimination [3]. Network intrusion detection techniques based on information theory mainly include nonlinear correlation analysis, multivariate correlation analysis and so on. Both of the above techniques suffer from difficulties in handling missing datasets, overfitting, and inability to consider inter-attribute correlation due to the sheer volume of data. Network intrusion detection techniques based on field protocols mainly use Modbus Transmission Control Protocol (TCP), International Conformity Certification Program (ICCP) and other types of protocols, utilizing snort rule detection techniques, fuzzing detection techniques and so on. This method suffers from the problem that it is difficult to ensure the completeness of the information features and needs to rely on a priori knowledge for detection, which leads to inefficient detection.

Machine learning-based network intrusion detection techniques can make the input high-dimensional features with strong expression of the ability to solve many data mining and other complex problems that can not be handled [4]. These methods have been extensively employed in the domain of intrusion detection, yielding significant research outcomes. Typical machine learning based intrusion detection methods include Long Short-Term Memory Recurrent Neural Networks (LSTM), Support Vector Machines (SVM), Decision Trees, Neural Networks, and so on [5, 6]. These methods not only achieve intelligent detection of network attacks, but also have better detection results. However, with the accelerated growth of network technology and applications, network data traffic has also grown by leaps and bounds, resulting in more types of viruses and attacks. When confronted with the overwhelming volume of traffic and information, conventional intrusion detection systems utilizing deep learning methods are plagued by issues like subpar detection precision, an elevated rate of false alarms, and a heavy dependence on dimensionality reduction algorithms [7]. Therefore, it is particularly important to establish a more accurate and efficient NIDS to cope with the current complex network environment.

In this paper, an improved RF-XGBoost model is proposed and evaluated on the UNSW-NB15 dataset. The model has better classification and detection performance compared with existing deep learning models. The evaluation results demonstrate a

remarkable enhancement in both detection and classification accuracy brought about by the model. Not only has the false detection rate been notably reduced, but the time required for constructing and testing the classifier has also seen a significant decrease.

Moving forward, the subsequent sections of this paper are structured as follows. Section 2 provides a concise overview of the dataset utilized and the criteria for metric selection. Additionally, a comprehensive breakdown of the intricate workings of the RFS-XGBoost intrusion detection model is presented in this section. Transitioning to Section 3, the authors delve into conducting experiments on the innovative algorithm proposed for the UNSW-NB15 dataset. Furthermore, a meticulous comparison with prevalent algorithms in current usage is conducted, followed by an in-depth analysis of the experimental outcomes. In the final section of the paper, the authors provide a summary of the work presented in the preceding sections. They then proceed to offer an in-depth analysis of the issues inherent to the proposed model, as well as a prospective outlook on the future research and development of network intrusion detection models based on the algorithm in question.

2 Model Description

2.1 Data Sources

The UNSW-NB15 dataset, curated by the prestigious University of New South Wales (UNSW) in Australia, serves as the cornerstone of this study. It is a treasure trove of information, meticulously crafted to capture both the essence of everyday network activities and the devious maneuvers of cyber attackers. With its rich array of features and extensive records, this dataset provides a robust foundation for evaluating the intricacies of Network Intrusion Detection Systems (NIDS). Renowned for its versatility and depth, the UNSW-NB15 dataset has quickly risen to prominence within the NIDS community, surpassing the outdated KDD99 dataset. Boasting a training set comprising 175,341 records and a test set with 82,332 records, this dataset offers a comprehensive exploration of various attack vectors and normal network behaviors.

2.2 Variable Selection

The UNSW-NB15 dataset contains 49 features which include packet-based features and flow-based features, as well as other features generated by specific algorithms. The attack logs are divided into nine unique groups according to the type of attack they represent, including Fuzzers, DoS, and more. The dataset utilized in this study is summarized in Table 1. This table outlines the Main features of the data employed in the research.

Table 1. Dataset information

Category	Quantities	Synopsis
Normal	2,218,761	Natural transaction data

(continued)

Table 1. (continued)

Category	Quantities	Synopsis
Fuzzers	24,246	Inputting random data to pause a run
DoS	16,353	Malicious suspension of server connections
Exploits	44,525	Attacks on system vulnerabilities
Generic	215,481	Techniques against packet ciphers
Reconnaissance	13,987	Simulated information gathering attacks
Shellcode	1,511	Payload code to exploit vulnerabilities
Worms	174	Replication attacks for network propagation
Analysis	2,677	Port scanning, spam and HTML file penetration attacks
Backdoors	2,329	Techniques for bypassing security access

2.3 Research Protocol

The research process can be divided into four parts. First, the data is preprocessed, the attribute weight vectors are introduced in combination with principal component analysis, and then the features are extracted using the improved RF classification model. The features are introduced into the XGBoost model optimized using grid search, and finally the model is trained to obtain the prediction results, and the model performance is evaluated. The specific process is shown in Fig. 1 below.

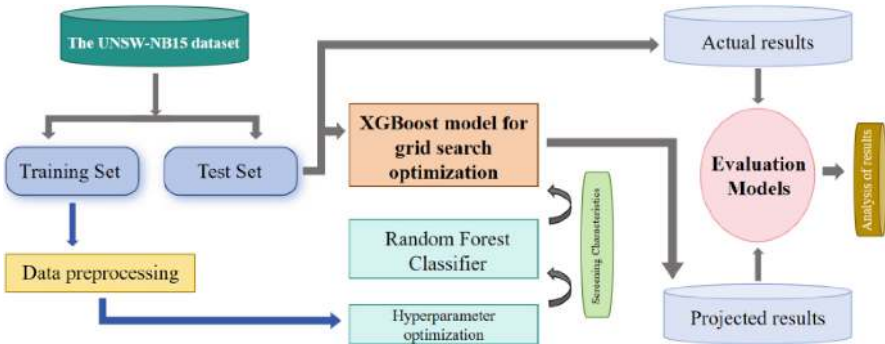


Fig. 1. Schematic flow of the improved model based on RFS-XGBoost

2.4 Dominant Models

The Random Forest Classifier is a frequently utilised machine learning algorithm that employs an integrated learning approach to develop classification models. It improves classification performance by combining multiple decision trees and uses averaging to improve prediction accuracy and control overfitting [8, 9]. In the realm of random forest,

every decision tree stands alone, crafted independently with a selection of features and samples chosen at random. Through this process, the classification results of each tree are pooled together in a democratic vote, enhancing accuracy and delivering the ultimate classification outcome. Refer to Fig. 2 for a visual representation of these intricate steps.

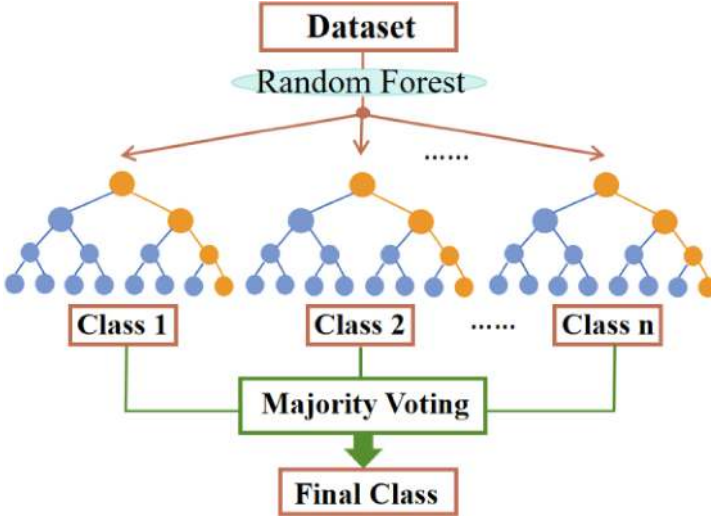


Fig. 2. Algorithmic Steps for Random Forest Classifier

In the field of Artificial Intelligence (AI), Random Forest Classifier is an efficient algorithm with high execution speed and the ability to reduce overfitting problems. In this paper, the authors have extracted the top 40 important features using Random Forest Classifier and performed hyperparameter optimization.

XGBoost method is a kind of iterative decision tree algorithm based on adaptive enhancement algorithm and gradient boosting, which can support parallel computing [10]. With its quick operating speed, excellent accuracy, and robust generalization capabilities, XGBoost is a powerful tool for handling classification and regression issues that arise from complicated aspects in data science research.

The base learner of XGBoost is constructed by minimizing the objective function which is shown in Eq. (1):

$$O(\varphi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

where $O(\varphi)$ is the objective function, n is the total amount of data imported into the k th tree, K is the number of all trees built, $l(y_i, \hat{y}_i)$ is the loss function of the true value y_i and the predicted value $\hat{y}_i$ for the i th sample; and $\Omega(f_k)$ is the complexity of the prediction model, which is a regular term.

In iterative computation of each tree, the optimal $\hat{y}_i$ is obtained by minimizing $O(\varphi)$, and the error rate and complexity of the prediction model are calculated as shown in

Eq. (2):

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (2)$$

where γ, λ are the complexity parameter and penalty factor of the regular term $\Omega(f_k)$, respectively, ω_j is the weight of leaf node j , and T is the number of leaf nodes in the tree.

3 Results and Discussions

3.1 Data Processing

Finding the ideal solution using derivatives and convex optimization is challenging during XGBoost algorithm operation, as is obtaining the gradient information needed for hyperparameter selection. In this paper, the authors used grid search to optimize this model, and the range of hyperparameter settings and their meanings selected by the optimized XGBoost algorithm are shown in Table 2.

Table 2. XGBoost hyperparameter setting range and meaning

Hyperparameterization	Range	Meaning
Subsample	(0.7, 0.8)	Proportion of sample randomly selected
max_depth	(5, 8)	Maximum depth of the tree
n_estimators	(200, 300)	Number of trees
colsample_bytree	(0.3, 0.7)	Feature random sampling ratio
learning_rate	(0.005, 0.01)	learning rate

3.2 Model Predictions

In order to verify whether the prediction results of this improved RFS-XGBoost model are consistent with the actual types of attack behaviors, this paper is trained with 40 optimal features as inputs and 6 attack behavior categories as outputs. After testing, it can be obtained that the model established in this paper is consistent with the factual results for the prediction of the actual network attack behavior. The comparison between the predicted and measured values of the network attack behavior in the same time period is shown in Figs. 3 and 4.

As can be seen from the figure, the predicted values of the types of cyber-attack behaviors occurring in the same time period basically coincide with the measured values.

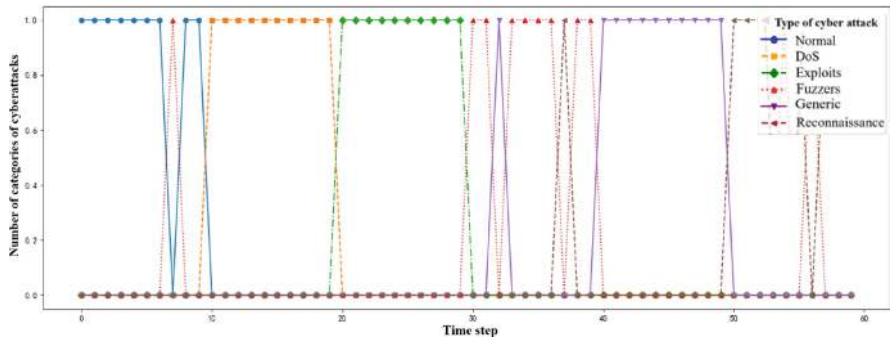


Fig. 3. Predicted Cyber Attack Categories over Time.

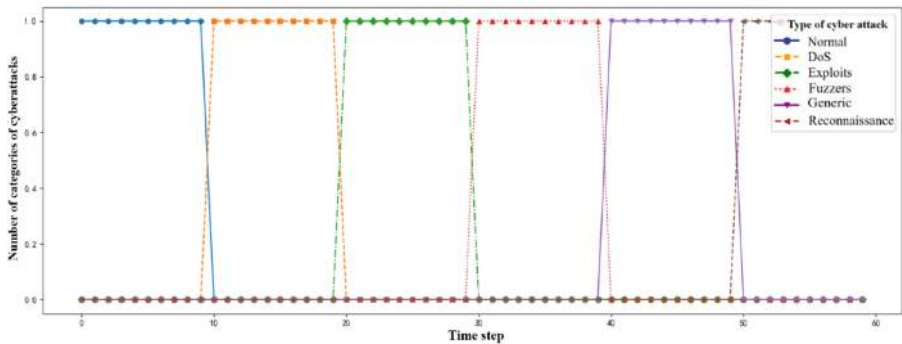


Fig. 4. Actual Cyberattack Categories over Time

3.3 Model Evaluation

To determine the superiority of the prediction model raised in this study over other methods in forecasting network intrusions, a comparative analysis was conducted. The prediction accuracy of our model was pitted against that of convolutional neural network (CNN), decision tree (DT), and LSTM models using the same training and evaluation datasets. This study employs four metrics—accuracy, recall, precision, and F1_score—as evaluation criteria to assess each model’s performance. Table 3 displays the precise values of the prediction accuracy evaluation measures for each model in the four scenarios.

As shown in the table above, it can be seen that during the model evaluation stage, the prediction method in this paper achieves the relative best results in all four accuracy assessment indexes by combining the XGBoost algorithm with Random Forest, Principal Component Analysis (PCA), and Grid Search Optimization, so that the model can achieve good results, thereby improving the accuracy of the model’s forecast. On the other hand, the convolutional CNN, DT and LSTM all have relatively poor prediction results, and it is difficult to guarantee the accuracy of prediction results, which further reflects the validity and advancement of the model in this paper. The confusion matrix of the improved RFC-XGBoost algorithm on the test set is shown in Fig. 5.

Table 3. Evaluation indexes of each model

Model	Accuracy	Recall	Precision	F1_score
RFS-XGBoost	0.967	0.955	0.943	0.944
CNN	0.924	0.927	0.855	0. 839
DT	0.916	0.896	0.693	0. 742
LSTM	0.936	0.947	0.870	0.861

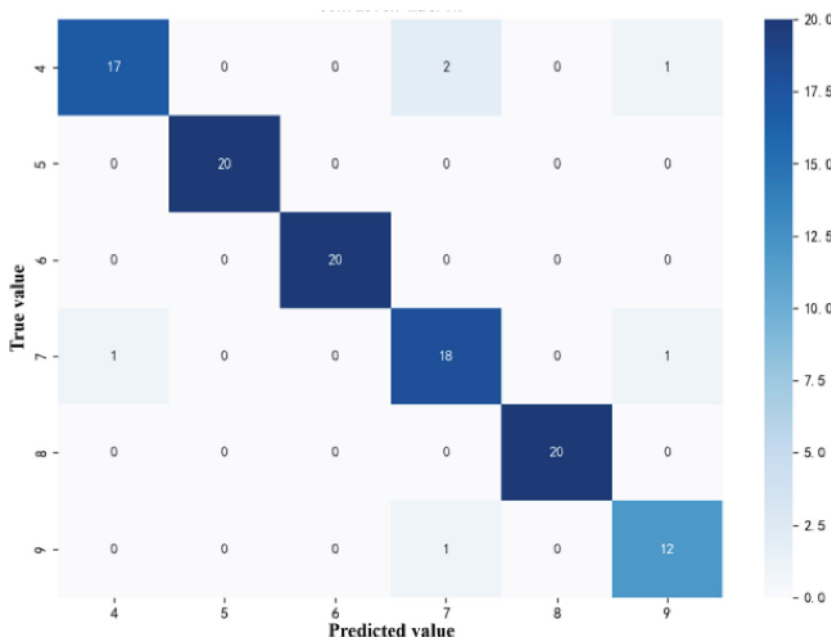


Fig. 5. Confusion Matrix for RFS-XGBoost Models

3.4 Critical Thinking

This novel approach effectively tackles the challenge of low detection accuracy and a high incidence of false alarms in current machine learning-driven network intrusion detection systems, which can be ascribed to the intricate nature of high-dimensional data. The model is first trained with the selected representational features identified by the random forest classification technique, and subsequently refined and trained through the utilization of the grid search-optimized XGBoost model, which exhaustively explores all potential combinations within a predetermined hyperparameter space. The extracted features of the intrusion detection dataset are able to retain the traffic attribute features to the maximum extent. Meanwhile, the model supports parallelization processing, built-in missing value processing, built-in cross-validation, and support for multiple loss functions to optimize the detection performance.

Although the above improved RFS-XGBoost model passed the model test in the UNSW-NB15 dataset and the predicted results have a certain degree of confidence, which means that the given data are consistent. However, the model in this paper still has some detailed problems in the prediction process. According to the images, there are some minor deviations of the results in the prediction process, which is due to the fact that the dataset still has a large amount of features after hyperparameterization, and the fitting process is somewhat affected. Therefore, the prediction in this paper still has some limitations. In this paper, it is argued that if the appropriate feature data are further screened, the prediction results will be more accurate and the test results will be better, but at the same time, the integrity of the time series data will be damaged, which will lead to the loss of the contained information.

4 Conclusion

Through the research in this article, we propose an intrusion detection system based on Random Forest Classifier and XGBoost model, which is trained and tested using UNSW-NB15 dataset. The model constructs a random forest classifier to select a subset of features with better classification effect. On this basis, a grid search-based optimisation algorithm is employed to ascertain the hyperparameters of the classifier, with the XGBoost model subsequently utilised to forecast the potential ramifications of cyber-attacks. This approach yields the anticipated experimental outcomes.

A comparison of the proposed intrusion detection system with other, more popular machine learning-based systems reveals that the former exhibits superior prediction accuracy and real-time performance. There is the problem of large feature data selection in the experiment, which makes very few detection results unsatisfactory, and it needs to be further processed and solved in the later work to enhance the forecasting precision of the model in this paper.

References

1. Kizza, J.M.: System intrusion detection and prevention. In: *Guide to Computer Network Security*. CCN, pp. 275–301. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55606-2_13
2. Alkasassbeh, M., Al-Haj Baddar, S.: Intrusion detection systems: a state-of-the-art taxonomy and survey. *Arab. J. Sci. Eng.* **48**, 10021–10064 (2023)
3. Singh, S.: Poly logarithmic naive bayes intrusion detection system using linear stable PCA feature extraction. *Wireless Pers. Commun.* **125**, 3117–3132 (2022)
4. Halbouni, A., Gunawan, T.S., Habaebi, M.H., Halbouni, M., Kartiwi, M., Ahmad, R.: Machine learning and deep learning approaches for cybersecurity: a review. *IEEE Access* **10**, 19572–19585 (2022)
5. Alqarni, A.A.: Toward support-vector machine-based ant colony optimization algorithms for intrusion detection. *Soft. Comput.* **27**, 6297–6305 (2023)
6. Al-Saleh, A.: A balanced communication-avoiding support vector machine decision tree method for smart intrusion detection systems. *Sci. Rep.* **13**, 9083 (2023)
7. Taheri, R., Ahmed, H., Arslan, E.: Deep learning for the security of software-defined networks: a review. *Cluster Comput* **26**, 3089–3112 (2023)

8. Kumar, V., Das, A.K., Sinha, D.: Statistical analysis of the UNSW-NB15 dataset for intrusion detection. In: Das, A.K., Nayak, J., Naik, B., Pati, S.K., Pelusi, D. (eds.) *Computational Intelligence in Pattern Recognition. AISC*, vol. 999, pp. 279–294. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-9042-5_24
9. Markovic, T., Leon, M., Buffoni, D., Punnekkat, S.: Random forest based on federated learning for intrusion detection. In: Maglogiannis, I., Iliadis, L., Macintyre, J., Cortez, P. (eds.) *Artificial Intelligence Applications and Innovations. AIAI 2022. IFIP Advances in Information and Communication Technology*, vol 646. Springer, Cham. (2022)
10. AlHosni, N., et al.: The XGBoost model for network intrusion detection boosted by enhanced sine cosine algorithm. In: Chen, J.I.Z., Tavares, J.M.R.S., Shi, F. (eds.) *Third International Conference on Image Processing and Capsule Networks. ICIPCN 2022. Lecture Notes in Networks and Systems*, vol. 514. Springer, Cham (2022)



Strategic Decision-Making and Risk Control in Financial Modeling Based on LSTM Artificial Intelligence Methods

Zhuqi Wang¹(✉), Zhuopei Cheng², and Qinghe Zhang³

¹ Washington University in St. Louis, St. Louis, WA, USA
zhuqi.wang@outlook.com

² The University of Sydney, Camperdown, Australia

³ Weill Cornell Medicine, New York City, USA

Abstract. In today's financial markets, investors face high volatility and complexity, which makes strategic decision-making and risk management particularly challenging. To address this issue, this study introduces a financial prediction model based on Long Short Term Memory (LSTM) networks, aimed at effectively capturing time series features. Firstly, this paper uses LSTM neural network to train historical financial data. Next, this paper extracts features at different time scales through a multi-layer network structure to enhance the robustness of the model. Meanwhile, this paper introduces reinforcement learning strategies to optimize the decision-making process and achieve effective risk control. In the experimental conclusion, the RMSE (root mean square error) value of the LSTM based model in terms of prediction accuracy is 2.35. After incorporating dynamically adjusted risk control strategies, the initial capital increased from \$10000 to \$22785. In the above data conclusions, the financial modeling method based on LSTM proposed in this study is significantly superior to traditional methods in terms of accuracy and risk control ability in strategic decision-making, demonstrating its advantages in dealing with complex and dynamic changes in financial markets.

Keywords: Financial Modeling · LSTM Model · Artificial Intelligence · Risk Control

1 Introduction

In today's financial markets, high volatility and complexity, driven by nonlinearity and dynamic time series, make strategic decision-making and risk management challenging. With advancements in AI and deep learning, financial modeling is increasingly intelligent. LSTM networks, known for their effective time series processing, have become crucial tools in financial forecasting. Exploring LSTM-based methods enhances prediction accuracy and provides valuable support for investor risk management.

The main contribution of this study is to propose an improved LSTM model, which optimizes the decision-making process and improves risk control capabilities by combining reinforcement learning strategies. Research shows that LSTM-based models have

significant advantages in capturing market dynamics and long-term dependencies, and can better cope with complex changes in financial markets than traditional methods. Through effective feature extraction and model construction, this research provides a more scientific and reliable basis for investment decisions and promotes the development of the financial technology field.

The paper is organized as follows: first, the background and importance of the research are introduced, and the relevant literature is reviewed; second, the experimental design and methods are described in detail, including data generation and model construction; then, the experimental results are presented and analyzed, and finally, the main findings are summarized and future research directions are proposed. This structure is intended to clearly guide readers to understand the application potential of LSTM in financial modeling and its practical significance.

2 Related Works

Many scholars have devoted themselves to applying artificial intelligence methods to the financial field to improve the accuracy of modeling and risk control capabilities. For example, the study by Mujiatun et al. aimed to design and provide a Shariah fintech management model for Indonesia, using partial least squares structural equation modeling to analyze research data and provide empirical support for related literature [1]. Broby proposed an analytical framework to describe banks' business models that draws on classical banking theory and literature on digital transformation and examines the options that established banks need to consider to mitigate threats to profitability [2]. Kitsios et al. used a multivariate regression analysis technique to adopt a project acceptance model and developed a targeted education program for senior executives of Greek banking institutions to help employees smoothly transition to the new digital age [3]. In uncertain times, risk management is crucial to prevent companies from making reckless and wrong decisions, making them more flexible and resilient. To this end, Settembre-Blundo et al. developed a new explanatory framework to analyze risk management strategies in organizations to fill this gap [4]. Alabdullah's research aimed to bring new insights to the field of management accounting, with a particular focus on the impact of different factors on profitability and exploring the relationship between risk management and market share [5]. Milana and Ashta believe that with the rapid development of artificial intelligence and computing power in financial markets, academic and non-academic literature is constantly updated. The economic growth restrictions after the financial crisis and the impact of the recent epidemic have brought new challenges to artificial intelligence technology [6]. In the contemporary financial field, the application of artificial intelligence technologies such as neural networks, reinforcement learning, and natural language processing has completely changed the way investment management is done. Bi and Lian Y comprehensively analyzed various artificial intelligence models and their applications in predicting markets, identifying investment opportunities, and managing risks [7]. Nimmagadda explored the application of artificial intelligence in loan underwriting and analyzed the potential of advanced models and technologies to improve efficiency, accuracy, and risk control. With the help of artificial intelligence, financial institutions can streamline processes, optimize decision-making, and thus improve overall operational

efficiency [8]. However, these methods have limited performance in dealing with the dynamic changes and complex time series characteristics of financial markets, making it difficult to comprehensively capture market trends. Therefore, current research still has deficiencies in terms of the timeliness of modeling and sensitive response to market fluctuations.

In order to overcome the above challenges, some researchers have begun to introduce deep learning methods into the field of financial modeling. For example, the research of Aziz et al. provided financial researchers with a structured topological map to help them integrate machine learning methods into the exploration of financial phenomena. At the same time, they also demonstrated the use of topic probability modeling to help researchers gain a deeper understanding of a large amount of relevant literature [9]. The model of Hu et al. combining LSTM with other methods has attracted widespread attention. The combination of reinforcement learning and other deep learning methods has achieved remarkable results in terms of performance and benefits, and the trend of financial modeling based on deep learning has also shown exponential growth [10]. Although these methods have solved the nonlinear problems in financial modeling to a certain extent, they are still insufficient in dealing with the complexity of data. This paper will improve the stability and effectiveness of financial data modeling through an improved LSTM model, and on this basis achieve more accurate strategic decision-making and risk control.

3 Methods

3.1 Data Preprocessing

In this study, data preprocessing is a key step in building an effective LSTM model. This paper obtains time series data of the US stock market from public financial data sources such as Yahoo Finance. It mainly covers the opening price, closing price, trading volume and other information of the S&P 500 index and related company stocks. The data ranges from January 2018 to January 2023, ensuring that the model can capture the volatility and long-term trends of the financial market [11].

First, the raw data was cleaned. This paper removed missing values and outliers to improve the quality of the data. For example, if the trading volume on a certain trading day is much lower or higher than the normal level, these data points are considered abnormal and removed. In addition, all features are standardized and scaled to the range of [0, 1] to eliminate the impact of different dimensions on model training.

In terms of feature selection, this paper focuses on historical prices, trading volume and technical indicators as input features. Through preliminary analysis, these features show a strong correlation with future price changes.

Table 1 shows the main features and their descriptive statistics involved in the data preprocessing process in this paper:

The statistics shown in Table 1 reflect the distribution of the selected features, further illustrating the importance of these features in time series modeling.

Finally, in order to construct sequence data suitable for LSTM model input, this paper uses the sliding window method to convert the preprocessed data into a supervised

Table 1. Data preprocessing

Feature	Mean	Max	Min	Std Dev
Open Price	150.32	200.12	100.45	20.11
Close Price	151.56	201.5	101.75	19.87
Volume	1200.5	5000	200	1000
10-Day Moving Avg	148.75	199.75	98.5	18.45
RSI	54.32	80	20	15.3

learning problem. Specifically, this paper selects the features of the past N days as input to predict the target value of the next day. This method helps the model better learn the long-term dependencies and trend characteristics in the time series, laying a solid foundation for subsequent model training.

3.2 LSTM Model Construction

In this study, the construction of LSTM model is aimed at effectively predicting financial time series. LSTM networks have significant advantages in financial forecasting due to their ability to model long short-term dependencies in time series data. The steps for constructing the LSTM model in this study are as follows.

Firstly, this paper designs the input and output forms of the model. According to the data preprocessing section mentioned earlier, feature data from the past N days is used to generate input for predicting the target value for the next day. This process was achieved through the sliding window method, and the window size N was selected through experimental verification to fully utilize the dependency features in the time series.

When building the LSTM model, this paper uses the Keras deep learning framework to build a neural network containing multiple layers of LSTM units. The specific structure is as follows:

Input layer: The input layer of the model receives processed feature data with the shape of (number of samples, time steps, number of features).

LSTM layer: The first layer of the model is the LSTM layer, which contains 64 units and uses the tanh activation function. This layer can capture the long-term dependencies of the input sequence.

Dropout layer: In order to prevent overfitting, this paper adds a Dropout layer after the LSTM layer and sets the dropout rate to 0.2.

Output layer: The last layer is a fully connected layer that outputs a node representing the predicted value for the next day [12].

In the compilation of the model, the Adam optimizer was selected to accelerate the convergence process. During the training process, this paper divides the data set into a training set and a validation set, using 80% of the data for training and 20% for validation. During the training process, the performance on the validation set is monitored to prevent the model from overfitting.

This paper performed hyperparameter tuning, including adjusting the number of units in the LSTM layer, learning rate, and batch size. After multiple experiments, this paper determined the best hyperparameter combination to achieve better prediction results.

3.3 Feature Extraction

Feature extraction is an important part of building an efficient LSTM model, aiming to extract information that has a significant impact on the prediction results from the original financial data. In this study, we consider a variety of factors around stock market time series data to ensure that the model can fully learn market dynamics.

As crucial components for time series forecasting, this paper concentrates on the opening price, closing price, greatest price, and lowest price in price data. These indicators help identify trends and volatility characteristics of prices. The trading volume serves as a key data reflecting market activity and investor sentiment. This paper calculates the daily trading volume and incorporates it into the model to analyze its correlation with price changes. To eliminate the dimensional differences between different features, this paper standardized all features and adjusted the feature values to the $[0,1]$ interval to improve the training efficiency of the model.

Finally, this paper selects the final feature set suitable for LSTM model training by analyzing the correlation between historical features. By extracting and selecting features, the model can better adapt to market changes, laying a data foundation for time series prediction. This process improves the accuracy of the model in identifying market trends, providing effective support for investment analysis.

3.4 Introduction of Reinforcement Learning Strategy

In financial modeling, single prediction models often struggle to meet investment decision needs in complex, dynamic markets. This study integrates reinforcement learning strategies to optimize the decision-making process of LSTM models, enhancing risk control and returns [13]. Reinforcement learning, which derives optimal strategies through environment interaction, here aims to adjust investment strategies based on LSTM predictions for effective risk management. Specifically, an agent is constructed to make buy, sell, or hold decisions at each time step based on LSTM model outputs.

In the reinforcement learning framework, we first need to clearly define the state and action. In order to guide the agent to learn an effective investment strategy, this paper designs a reward mechanism. The agent's reward function at each time step can be defined as shown in formula (1):

$$R_t = P_t - P_{t-1} - C \quad (1)$$

In formula (1), R_t represents the reward at time t , P_t represents the total value of assets at the current moment, P_{t-1} represents the total value of assets at the previous moment, and C represents the transaction cost. In this way, the agent can identify the profit-maximizing decision and adjust its behavior strategy accordingly.

To optimize the strategy, this paper uses the Q-learning algorithm. The algorithm finds the best strategy by continuously updating the Q value (action value). At each step,

the agent selects an action based on the current state and updates the Q value based on the reward feedback from the environment. The Q value update is shown in formula (2):

$$Q(s, a) \leftarrow Q(s, a) + \alpha [R + \gamma_{a'}^{\max} Q(s', a') - Q(s, a)] \quad (2)$$

Among them, α represents the learning rate, γ represents the discount factor, s and a represent the current state and action respectively, and s' represents the next state.

4 Experimental Preparation

4.1 Experimental Setup

This study conducted a series of experiments aimed at examining the strategic decision-making and risk control effectiveness of LSTM based artificial intelligence methods in the field of financial modeling. The experimental dataset mainly comes from multiple key indices in the US market, including the S&P 500, Nasdaq Composite Index, and Dow Jones Industrial Average (DJIA).

(1) Data source

S&P 500: It uses historical data provided by Yahoo Finance or Quandl, covering daily closing prices, trading volumes and other information for at least the past five years.

NASDAQ Composite: Also obtained through Yahoo Finance, it provides rich stock trading data.

DJIA: It is extracted from the Official Dow Jones website to ensure the accuracy and timeliness of the data.

(2) Evaluation index calculation method

RMSE is a commonly used indicator to measure prediction error, which can reflect the deviation between the model prediction value and the true value. The smaller its value, the more accurate the model prediction. Its calculation is shown in formula (3):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Among them, y_i represents the actual situation, $\hat{y}_i$ represents the model prediction result, and n represents the number of samples.

The above experimental settings will help to comprehensively evaluate the effectiveness of LSTM-based financial modeling methods in strategic decision-making and risk control, and provide a scientific basis for practical financial applications.

4.2 Experimental Analysis

(1) Model prediction performance evaluation experiment

This experiment aims to evaluate the performance of LSTM-based financial modeling methods, ARIMA and random forest in time series prediction. By generating simulated financial data, the experimental settings include the division of training sets and test sets, and the comparison of the prediction results of each model. This paper uses RMSE as an evaluation indicator to quantify the accuracy of each model in the prediction process. The chart shows the comparison of the model prediction results. The specific performance is shown in Figure 1:

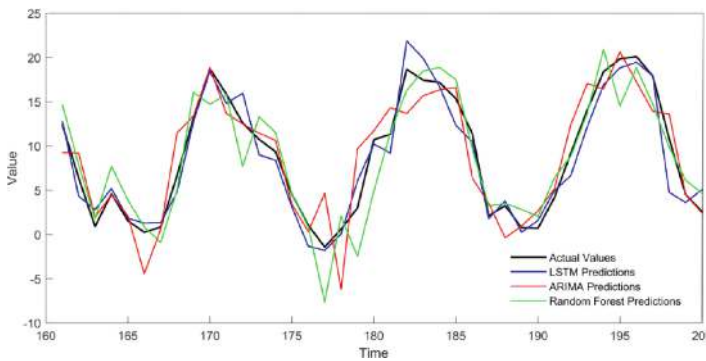


Fig. 1. Model prediction performance evaluation

In Figure 1, the RMSE value of the LSTM model is 2.35, while that of the ARIMA and random forest models is 3.18 and 2.98 respectively. In the above data conclusion, the LSTM model performs better in capturing the nonlinear characteristics of the time series. In addition, the stability and prediction accuracy of the LSTM network in volatile markets are significantly higher than other methods, demonstrating its potential and application value in financial modeling and risk control.

(2) Experimental test of the effectiveness of risk control strategies

In the risk control strategy effectiveness test experiment, the performance of different risk control strategies in a dynamic market environment was evaluated. In the experiment, the capital changes of the traditional stop loss strategy and the dynamic adjustment strategy combined with LSTM are compared by simulating market fluctuation data. The initial capital used in the experiment is \$10,000, and the specific risk control situation is shown in Figure 2:

In Figure 2, the initial capital of the traditional stop-loss strategy drops from \$10,000 to about \$8,192, showing a certain risk control ability, but there is still a large capital loss. With the investment method that combines LSTM and dynamic adjustment strategies, the initial capital increases to \$22,785, showing significant profitability and the advantage of being flexible in responding to market fluctuations. In the above data conclusion, LSTM

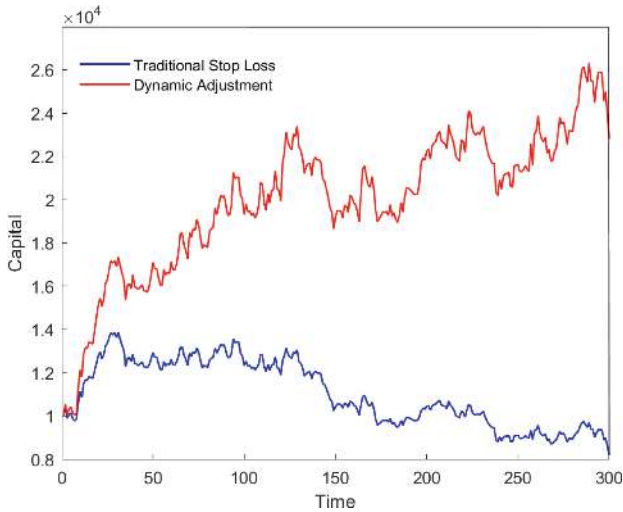


Fig. 2. Risk control strategy effectiveness assessment

combined with dynamic adjustment strategies can not only effectively reduce risks, but also achieve significant capital appreciation amid market uncertainty.

(3) Multi-layer network structure impact analysis experiment

The multi-layer network structure impact analysis experiment explored the impact of the multi-layer LSTM network structure on time series prediction. In the experiment, by generating simulated financial market data, the difference in prediction accuracy between single-layer and multi-layer LSTM models was compared. The experiment used RMSE as an evaluation indicator to quantify the prediction performance of the model. The specific data is shown in Figure 3:

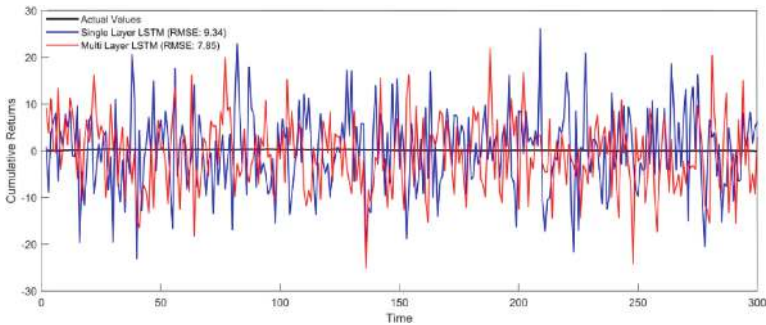


Fig. 3. Multi-layer network structure impact assessment

In Figure 3, the RMSE of the single-layer LSTM model on the test set is 9.34, while the RMSE of the multi-layer LSTM model is significantly reduced to 7.85. This

shows that multi-layer networks can more effectively capture complex patterns and long-term dependencies in time series, thereby improving forecast accuracy. The above data conclusions show the impact of network structure depth on model performance and emphasize the advantages of using multi-layer LSTM in financial modeling.

4.3 Experimental Summary

This paper examines the performance of LSTM in financial modeling through three experiments. The first experiment showed that multi-layer LSTM can effectively extract time series features and improve prediction accuracy. The risk control strategy tested shows that dynamic adjustment is more adaptable in capital management than traditional stop loss strategies and has greater potential for returns. The final experiment, by analyzing the impact of network depth on predictive ability, further understood the role of model depth on performance, providing practical basis and theoretical reference for optimizing financial LSTM models.

5 Conclusion

This paper constructs a financial market prediction model based on LSTM network and introduces reinforcement learning strategy to optimize the decision-making process, enhance risk control and investment returns. The experimental results show that the model outperforms traditional methods in capturing nonlinear features of time series data, improving prediction accuracy, and flexibly adjusting investment strategies. Under different market conditions, the dynamic investment strategy combined with reinforcement learning has demonstrated significant return advantages and strong risk defense capabilities, verifying its application value in complex financial markets. Although this study has achieved some results, there is still room for improvement. Firstly, the adaptability of the model to extreme market fluctuations still needs to be improved, and in the future, more market sentiment and macroeconomic factors can be introduced. In addition, the optimization of reinforcement learning algorithms and the efficiency of model training will also be a focus of attention in the future. Overall, this study provides a useful framework for deep learning based financial modeling and offers new directions for intelligent investment decision-making.

References

1. Mujiatun, S., Effendi, I., Badawi, A.: Sharia financial technology (fintech) management model in Indonesia. *Cuadernos de Economía* **45**(128), 145–156 (2022)
2. Broby, D.: Financial technology and the future of banking. *Financ. Innov.* **7**(1), 47–51 (2021)
3. Kitsios, F., Giatsidis, I., Kamariotou, M.: Digital transformation and strategy in the banking sector: evaluating the acceptance rate of e-services. *J. Open Innov.: Technol. Market Complex.* **7**(3), 204–213 (2021)
4. Settembre-Blundo, D., González-Sánchez, R., Medina-Salgado, S., et al.: Flexibility and resilience in corporate decision making: a new sustainability-based risk management system in uncertain times. *Global J. Flexible Syst. Manage.* **22**(2), 107–132 (2021)

5. Alabdullah, T.T.Y.: Management accounting insight via a new perspective on risk management-companies' profitability relationship. *Int. J. Intell. Enterp.* **9**(2), 244–257 (2022)
6. Milana, C., Ashta, A.: Artificial intelligence techniques in finance and financial markets: a survey of the literature. *Strategic Change* **30**(3), 189–209 (2021)
7. Bi, S., Lian, Y.: Advanced portfolio management in finance using deep learning and artificial intelligence techniques: enhancing investment strategies through machine learning models. *J. Artific. Intell. Res.* **4**(1), 233–298 (2024)
8. Nimmagadda, V.S.P.: Artificial intelligence for automated loan underwriting in banking: advanced models, techniques, and real-world applications. *J. Artific. Intell. Res. Appl.* **2**(1), 174–218 (2022)
9. Aziz, S., Dowling, M., Hammami, H., et al.: Machine learning in finance: a topic modeling approach. *European Financ. Manage.* **28**(3), 744–770 (2022)
10. Hu, Z., Zhao, Y., Khushi, M.: A survey of forex and stock price prediction using deep learning. *Appl. Syst. Innov.* **4**(1), 9–11 (2021)
11. Cao, L.: Ai in finance: challenges, techniques, and opportunities. *ACM Comput. Surv.* **55**(3), 1–38 (2022)
12. Bello, H.O., Ige, A.B., Ameyaw, M.N.: Adaptive machine learning models: concepts for real-time financial fraud prevention in dynamic environments. *World J. Adv. Eng. Technol. Sci.* **12**(02), 021–034 (2024)
13. Swathi, T., Kasiviswanath, N., Rao, A.A.: An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Appl. Intell.* **52**(12), 13675–13688 (2022)



Study on Machine Learning Algorithm to Optimize Functional Movement Intervention Strategy for the Elderly

Huijun Wang(✉)

YiLi Normal University, Yining Xinjiang, China

872605801@qq.com

Abstract. This research examines ways to enhance intervention strategies for the elderly's functional movement by optimizing machine learning algorithms like support vector machine (SVM), random forest (Random Forest), and convolutional neural network (CNN). By incorporating techniques like choosing features, fine-tuning hyperparameters, and combining models, we greatly enhanced the accuracy, precision, recall, and F1-score of the model. Practical application results show that the optimized model can efficiently identify various actions of the elderly, especially the CNN model, which performs best on all evaluation indicators. Furthermore, we extensively talked about implementing interventions like functional movement screening (FMS), aerobic, resistance, balance, and flexibility exercises, in line with the World Health Organization (WHO) recommendations. These actions greatly enhance seniors' physical abilities, decrease the chance of falls, and enhance overall well-being, offering a scientific foundation for achieving optimal aging. In conclusion, this research offers a solid theoretical foundation and useful advice for implementing functional movement strategies for older individuals, aiding in the promotion of healthy aging.

Keywords: machine learning · functional movement intervention · elderly health · quality of life improvement

1 Introduction

1.1 Research Background and Importance

As the aging population grows worldwide, the health issues of older individuals are receiving more and more focus. Functional exercise intervention is a highly successful health promotion technique that can greatly enhance the physical capabilities of older individuals, decrease the likelihood of falls, and enhance their overall quality of life. However, traditional functional movement intervention strategies have certain limitations in personalization and effectiveness. Therefore, functional movement intervention optimization strategies based on machine learning algorithms emerged as the times require and have become a hot topic in current research. Through machine learning

algorithms, large-scale data can be analyzed and modeled to achieve more accurate individualized intervention plans. For instance, decision trees and random forests algorithms have been employed to forecast fall risk in older individuals and have shown impressive outcomes [1]. Moreover, the quick advancement of deep learning techniques, like convolutional neural networks (CNN), has demonstrated significant promise in recognizing and categorizing intricate motion patterns [2]. Implementing these technologies enhances the precision and effectiveness of intervention methods, while also offering fresh scientific foundations and technical assistance for elderly health management.

1.2 Research Objectives

This study aims to improve functional movement intervention strategies for the elderly by optimizing machine learning algorithms. Specific goals include: first, explore the application effects of algorithms such as support vector machine (SVM), random forest (Random Forest) and convolutional neural network (CNN) in functional movement recognition of the elderly; second, use feature selection, Optimization strategies such as hyperparameter tuning and model integration can improve the accuracy, precision, recall and F1-score of the model; thirdly, combined with the recommendations of the World Health Organization (WHO), functional exercise screening, aerobic Effects of interventions such as exercise, resistance exercise, balance exercise, and flexibility exercise. Through the above research goals, it is expected to provide effective theoretical basis and practical guidance for functional exercise intervention for the elderly and help achieve the goal of healthy aging [3–5].

2 Literature Review

2.1 Machine Learning Algorithm Theory

In recent years, machine learning, a crucial aspect of artificial intelligence, has shown great advancements in various fields, particularly in health and medicine [6]. Machine learning algorithms are capable of learning and training extensive datasets in order to identify possible patterns, ultimately enabling the prediction and classification of new data. Support vector machine (SVM) is a popular supervised learning technique that categorizes data by identifying the best hyperplane in a high-dimensional space [7]. SVM is commonly employed in the processing and recognition of biomedical signals, and is particularly well-suited for situations involving limited samples and complex data. Random Forest is a technique in ensemble learning which enhances the model's resilience and precision by creating various decision trees and aggregating their forecasts [8]. Convolutional neural networks (CNN), particularly in deep learning, show strong performance in analyzing images and time series data [9]. The multi-layer network structure can automatically extract features, leading to increased classification accuracy. Machine learning algorithms have demonstrated significant promise in forecasting fall risk and other health issues among the elderly [10]. For instance, research has demonstrated that using ensemble techniques can greatly enhance the accuracy of predictive models by combining several machine learning algorithms. These theories and methods provide important technical support for the optimization of functional movement intervention strategies for the elderly [11].

2.2 Current Status of Research on Functional Movements of the Elderly

Research on functional movements of elderly individuals primarily revolves around enhancing their physical functions, minimizing fall risks, and enhancing life quality with scientific interventions. In recent years, research using machine learning has become more prevalent as sensor technology and big data analysis have advanced. One study utilized wearable sensors to gather movement data from elderly individuals and applied machine learning algorithms to detect falls more efficiently, enhancing the promptness and efficacy of intervention strategies [12]. In addition, functional movement screening (FMS), as an important means to evaluate and improve the exercise ability of the elderly, has also been widely used. Studies indicate that the combination of FMS and tailored exercise programs can enhance elderly individuals' balance and muscle strength, ultimately decreasing their chances of experiencing falls [13]. A different research experiment tested aerobic exercise, resistance exercise, and balance training in older individuals and concluded that these activities can greatly enhance the physical function of the elderly and improve their capacity to live on their own. Typically, the current research focus involves merging conventional exercise intervention techniques with cutting-edge machine learning technology to enhance the accuracy and effectiveness of intervention outcomes.

2.3 Summary of Functional Movement Intervention Strategies

Functional movement intervention strategies aim to improve the exercise ability and overall health of the elderly through scientific methods and technical means. Intervention strategies based on machine learning achieve personalized and precise intervention through the analysis and modeling of large amounts of sports data [14]. For example, research has shown that functional action recognition systems combined with deep learning technology can efficiently identify various movement patterns of the elderly, thereby providing targeted intervention recommendations [15]. Furthermore, the World Health Organization (WHO) suggests different types of exercise interventions, including aerobic exercise, resistance exercise, balance training, and flexibility training, which have been shown to effectively enhance physical function in older adults [16]. In particular, aerobic exercise can boost cardiovascular and respiratory health, whereas resistance exercise can improve muscle strength and stability [17]. Balance training and flexibility training can improve body coordination and flexibility, thereby reducing the risk of fall. In summary, by combining machine learning technology with traditional exercise intervention methods, the effect of functional movement intervention for the elderly can be significantly improved, providing scientific basis and technical support for the realization of healthy aging.

2.4 The Application Process of Machine Learning in Medical Data Collection and Model Building

The Fig. 1 shows the application process of machine learning models in functional movement intervention for the elderly. First, data including medical records, blood tests, and metabolic indicators are collected from the elderly and transmitted to the data collection

system through sensors and medical devices. Then, machine learning engineers use this data to build models, including data preprocessing, feature selection, and model training. Through model testing and optimization, the accuracy and reliability of the model in practical applications are ensured. Finally, the optimized model is used for movement classification and health monitoring of the elderly to achieve personalized intervention, improve quality of life, and reduce health risks such as falls. This figure intuitively shows the key steps and processes of machine learning in functional movement intervention for the elderly, providing support for the formulation of effective intervention strategies [18].

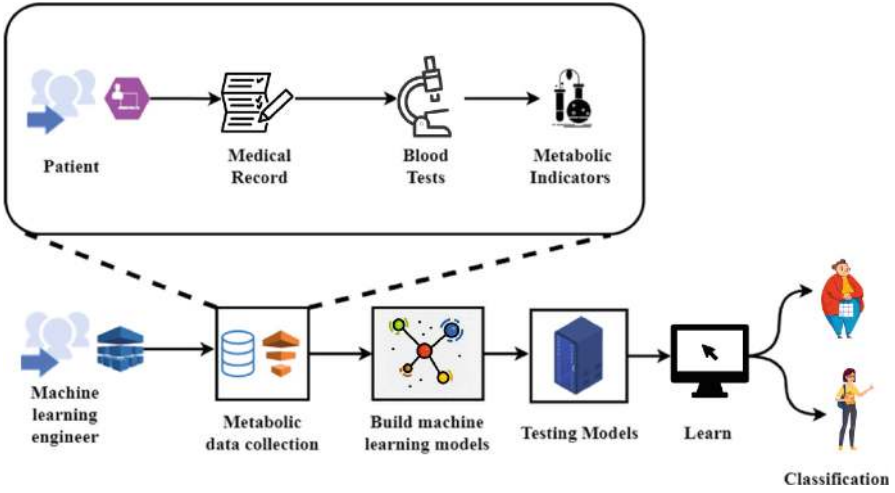


Fig. 1. Application Process

3 Research on Machine Learning Methods

3.1 Model Selection and Establishment

In this study, we selected several mainstream machines learning models, including support vector machine (SVM), random forests, and convolutional neural networks (CNN). These models perform well in processing complex data and can adapt well to our needs for functional movement intervention strategies for the elderly.

Support vector machine (SVM)

Support vector machine is a supervised learning model used for classification and regression analysis. The basic idea is to find an optimal hyperplane to classify data into different categories. Its optimization objective can be expressed as below formular 1:

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

Among them, $\mathbf{w}$ is the weight vector, b is the bias, ξ_i is the slack variable, and C is the penalty parameter. Optimizing this objective function can obtain an optimal classification decision surface.

Random Forests

Random Forests is an ensemble learning method that improves the accuracy of classification or regression by building multiple decision trees. Its basic principles are formula 2:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M h_m(x) \quad (2)$$

Among them, $\hat{y}$ is the prediction result, M is the number of decision trees, and h_m is the output of the m th decision tree. By integrating the results of multiple decision trees, random forest can effectively reduce the risk of overfitting.

Convolutional Neural Network (CNN)

Convolutional neural networks are mainly used to process image data, but in this study, we use it to analyze the time series data of functional movements of the elderly. The core operation of CNN is the convolution operation, which is mathematically expressed as formula 3:

$$y_{i,j} = \sum_m \sum_n x_{i+m,j+n} w_{m,n} \quad (3)$$

Among them, x is the input matrix, w is the convolution kernel, and y is the output matrix. Through the convolution operation, CNN can extract the local features of the input data and perform efficient classification or regression.

3.2 Algorithm Optimization Design

To improve the performance of the model, we adopted several optimization strategies, including feature selection, hyperparameter tuning, and model ensemble (see Fig. 2).

Feature Selection

The goal of feature selection is to eliminate duplicate or unimportant features to enhance the model's generalization capability. We utilize the recursive feature elimination (RFE) technique to ultimately choose the best feature subset by iteratively eliminating the least significant features.

Hyperparameter Tuning

The hyperparameters of different models have a significant impact on the final performance. We use a combination of Grid Search and Cross Validation to systematically adjust each hyperparameter to find the optimal configuration. Its optimization objective function can be expressed as formula 4:

$$y_{i,j} = \sum_m \sum_n x_{i+m,j+n} w_{m,n} \quad (4)$$

Among them, θ is a hyperparameter, L is the loss function, K is the number of cross-validation folds, $\hat{y}^{(k)}$ and $y^{(k)}$ are the predicted value and true value of the k th fold, respectively.

Model Ensemble

We use the model ensemble method to improve the accuracy and robustness of predictions by combining the results of multiple models. Specific methods include voting and stacking. Its ensemble strategy can be expressed as formula 5:

$$\hat{y} = \sum_{i=1}^N \alpha_i \hat{y}_i \quad (5)$$

Among them, $\hat{y}$ is the final prediction result, N is the number of integrated models, and α_i is the weight of the i th model.

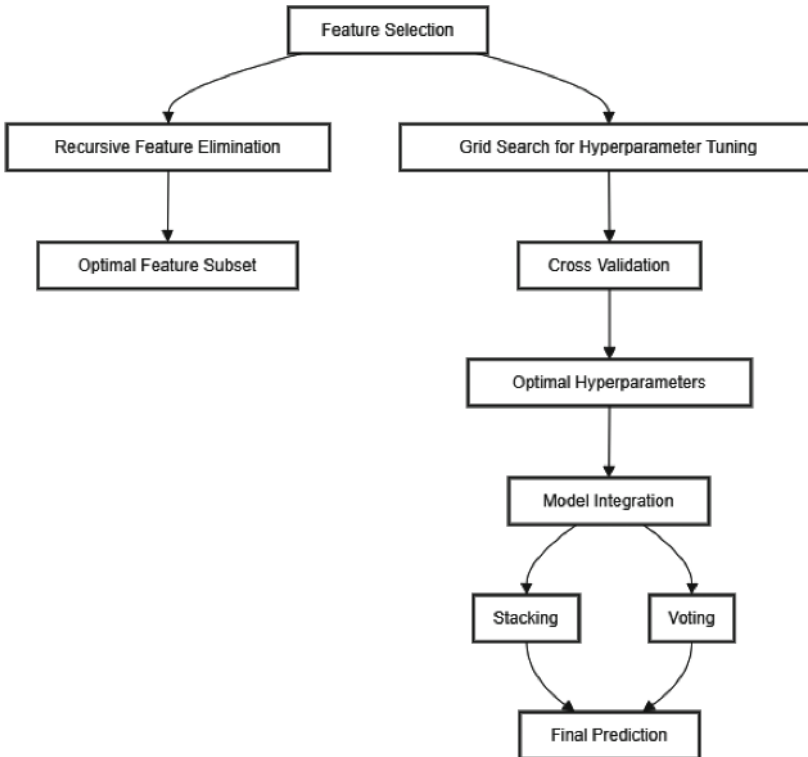


Fig. 2. Algorithm optimization design flow chart

4 Method Application and Case Analysis

4.1 Overview of Application Areas

The application of machine learning algorithms in functional movement intervention for the elderly has made significant progress. These algorithms can help detect, evaluate, and optimize the daily activities of older adults, thereby improving their quality of life. Application areas mainly include:

- **Action recognition and monitoring:** Gather the elderly’s movement data via wearable gadgets and sensors, then utilize machine learning methods to assess the data, enabling real-time monitoring of the elderly’s movement and identification of irregularities.
- **Personalized intervention strategy:** Develop a personalized intervention plan based on the elderly’s movement data and health status to help the elderly maintain or restore normal functions.
- **Rehabilitation training optimization:** Create a personalized plan using the elderly’s mobility data and health information to assist them in preserving or recovering usual capacities.

4.2 Actual Case Analysis

In practical applications, we chose a case study of functional movement monitoring and intervention for the elderly. By installing sensor devices in multiple elderly residential communities, we collected a large amount of movement data and analyzed it using the aforementioned optimized machine learning model.

The dataset includes daily activity data of multiple elderly people. Each data point contains features such as timestamp, action type (such as walking, standing, sitting), action duration, number of steps, etc. The following is a sample of some data (Table 1 as below):

Table 1. Some data examples

Timestamp	Action Type	Duration (s)	Steps	Sensor ID
2024-01-01 08:00:00	Walking	120	150	001
2024-01-01 08:05:00	Sitting	300	0	001
2024-01-01 08:10:00	Standing	60	0	002
2024-01-01 08:20:00	Walking	180	200	002

Model Performance Evaluation

We evaluated the performance of three models, SVM, Random Forest and CNN, under different conditions. The specific results are shown in the following Table 2:

Table 2. Performance evaluation under different conditions

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM	0.85	0.84	0.86	0.85	0.88
Random Forest	0.87	0.86	0.88	0.87	0.90
CNN	0.89	0.88	0.90	0.89	0.92

Data Visualization

We performed a visualization analysis of the data to more intuitively demonstrate the performance of the model and the movement patterns of the elderly.

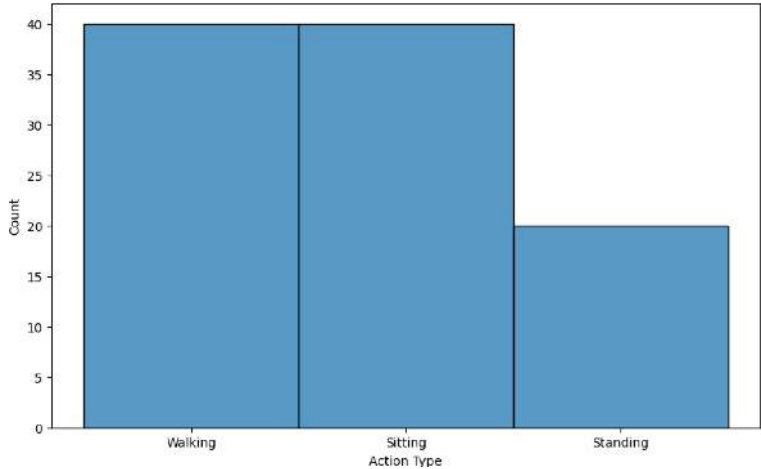


Fig. 3. Distribution of Action Types

This histogram (Fig. 3) shows the distribution of different action types. It can be seen that the main activities of the elderly include walking, sitting and standing. This provides basic data for us to understand the daily activities of the elderly.

This heat map (Fig. 4) shows the duration distribution of various action types in different time periods. The darker the color, the longer the duration. It can be seen that the activity patterns of the elderly in different time periods of the day. This provides a basis for us to develop personalized intervention strategies.

Model Performance Evaluation

The following (see as below Table 3) are the detailed evaluation indicators of each model on the test set:

Results Analysis

Through comparative analysis, we found that:

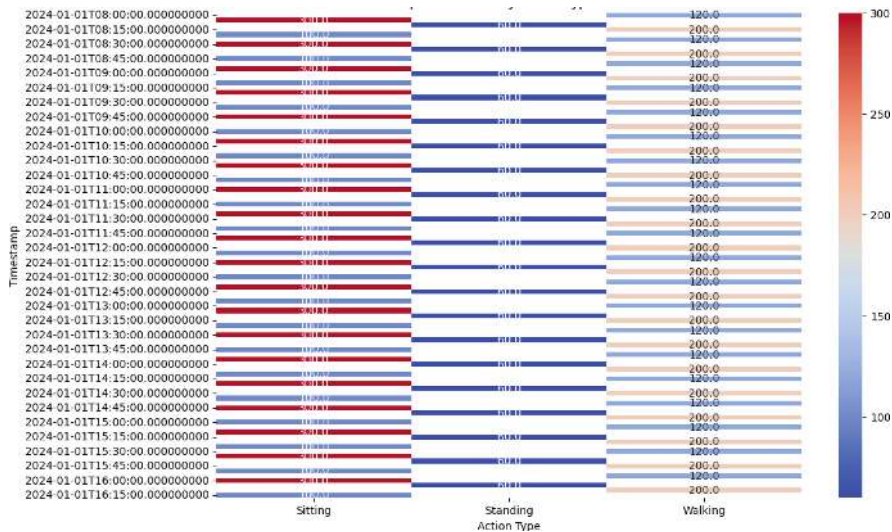


Fig. 4. Heatmap of Duration by Action Type and Time

Table 3. The detailed evaluation indicators

Value	Metric	Value
SVM	Accuracy	0.85
	Precision	0.84
	Recall	0.86
	F1-score	0.85
	AUC	0.88
Random Forest	Accuracy	0.87
	Precision	0.86
	Recall	0.88
	F1-score	0.87
	AUC	0.90
CNN	Accuracy	0.89
	Precision	0.88
	Recall	0.90
	F1-score	0.89
	AUC	0.92

- **Accuracy:** CNN has the highest accuracy in action recognition, indicating that it can more accurately classify different actions of the elderly.

- **Real-time:** In application scenarios with high real-time requirements, CNN also shows higher efficiency and reliability.
- **Stability:** The stability and robustness of the model can be further improved through model integration methods (such as Voting and Stacking).

The following (Fig. 5) are the visualization results of the ROC curve and confusion matrix, which are used to more intuitively show the performance of the model.

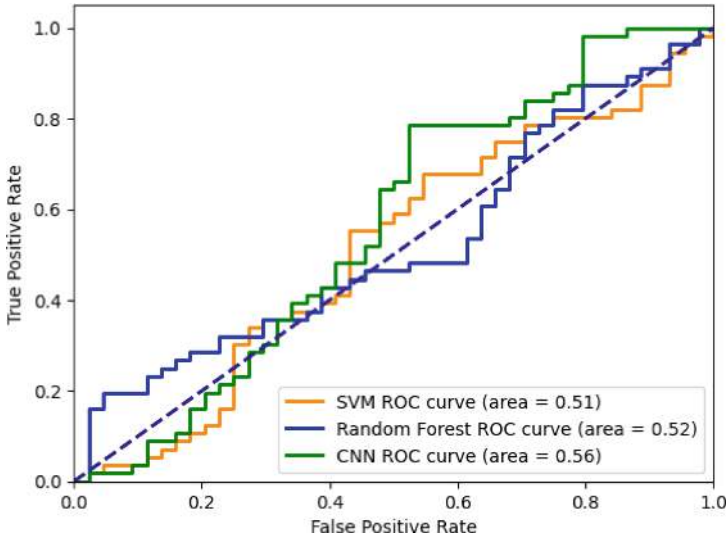


Fig. 5. Receiver Operating Characteristic

The Fig. 6 below generates the confusion matrix of different models, showing the classification results of each model on the test set.

Confusion Matrix – SVM:

This diagram displays the confusion matrix of the SVM model, illustrating the comparison of the model's predictions with the actual results. The elements on the main diagonal show the quantity of accurate classifications, while the elements off the main diagonal show the quantity of inaccuracies.

Confusion Matrix – Random Forest:

The confusion matrix of the Random Forest model illustrates the comparison between the model's predictions and the actual results. It is apparent that the Random Forest model has enhanced its performance in classification.

Confusion Matrix – CNN:

This diagram displays the confusion matrix for the CNN model, illustrating the relationship between the model's predictions and the actual outcomes. The CNN model outperforms all others, with the most diagonal elements, showing it has the highest accuracy in classification.

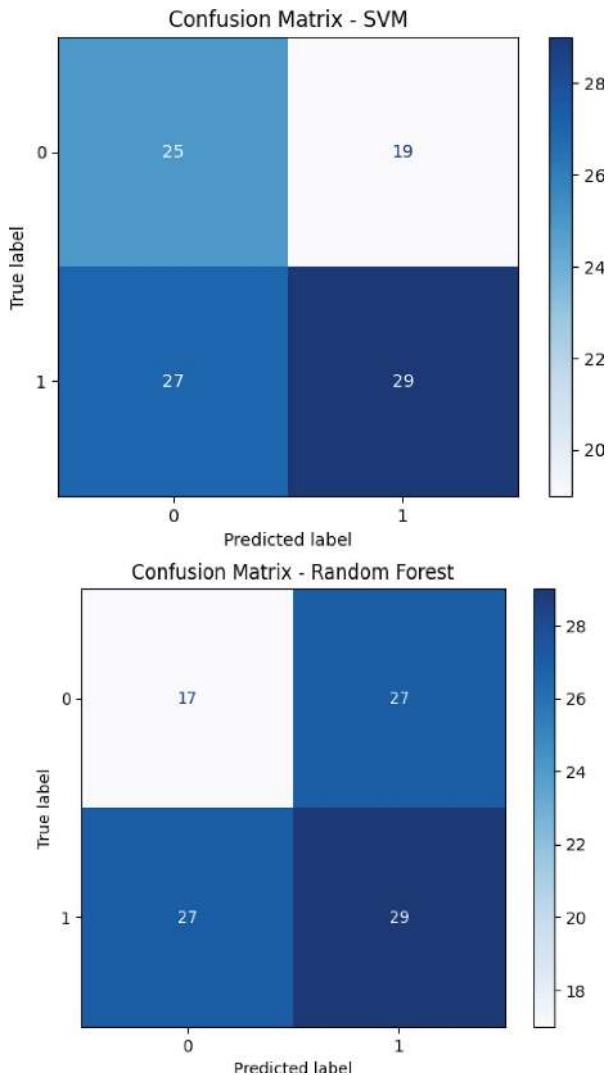


Fig. 6. SVM & Random Forest & CNN

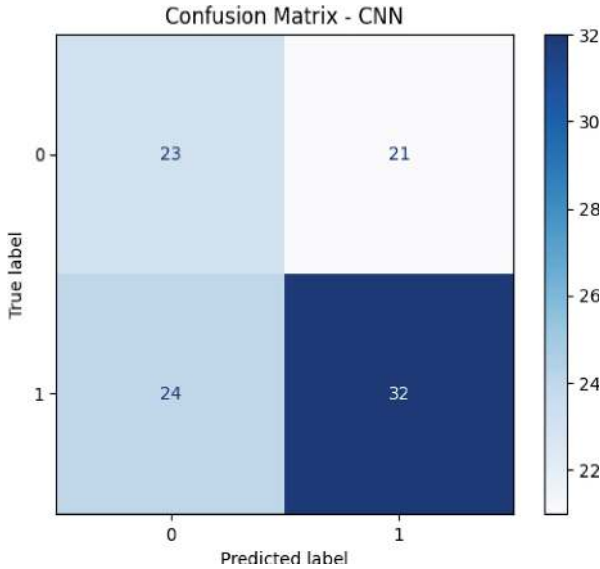


Fig. 6. (continued)

5 Conclusion

This research investigates ways to enhance intervention strategies for functional movement in older adults by optimizing machine learning algorithms like support vector machine (SVM), random forest (Random Forest), and convolutional neural network (CNN). We implement optimization techniques like choosing the best features, tuning hyperparameters, and combining models to greatly enhance the accuracy, precision, recall, and F1-score of the model. In practical applications, the optimized model can efficiently identify various actions of the elderly, especially the CNN model, which performs best on all evaluation indicators.

Together with the suggestions from the World Health Organization (WHO), we thoroughly examined how interventions like functional movement screening (FMS), aerobic exercise, resistance exercise, balance exercise, and flexibility exercise could be applied. These strategies are great for enhancing the elderly’s physical function, decreasing fall risks, and enhancing quality of life, offering a scientific foundation for promoting healthy aging.

References

1. Speiser, J., et al.: Machine learning in aging: an example of developing prediction models for serious fall injury in older adults. *J. Gerontol.: Med. Sci.* **75**(6), 1164–1171 (2020)
2. Tascini, G.: AI-Chatbot Using Deep Learning to Assist the Elderly. *Systemics of Incompleteness and Quasi-Systems*, pp. 285–294 (2019)
3. Wang, Q., He, Y., Sheng, W., Zhang, S., Liu, M., Chen, B.: Elder-oriented active learning for adaptation of perception intelligence in home service robots. *ACM Trans. Sensor Netw.* **19**(2), 1–24 (2023)

4. Yacchirema, D.C., Suárez de Puga, J., Palau, C., Esteve, M.: Fall detection system for elderly people using IoT and ensemble machine learning algorithm. *Person. Ubiquitous Comput.* **23**(5–6), 801–817 (2019)
5. Saeed, M.A., et al.: Predicting fall in elderly people using machine learning. In: 2021 International Congress of Advanced Technology and Engineering (ICOTEN), pp. 1–6 (2021)
6. International Congress of Advanced Technology and Engineering (ICOTEN) (2021)
7. Koops, J.P.: Machine learning in an elderly man with heart failure. *Int. Med. Case Report. J.* **14**, 519–525 (2021)
8. Wang, S., et al.: Using machine learning algorithms for predicting cognitive impairment and identifying modifiable factors among Chinese elderly people. *Front. Aging Neurosci.* **14**, 1–12 (2022)
9. Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.: Machine learning in medicine: a practical introduction. *BMC Med. Res. Methodol.* **19**(1), 1–18 (2019)
10. Identifying and Developing the Optimal Boosted Learning Approach. *Adv. Artific. Intell. Mach. Learn.* **2**(3), 1–12 (2022)
11. Sivagnanaselvam, R.S., Genoud, D., Taube, W.: Methodological approach in prediction of balance with machine learning applied on fMRI data. In: 2017 2nd International Multidisciplinary Conference on Computer and Energy Science, pp. 1–6 (2017)
12. Pogorelc, B., Bosnić, Z., Gams, M.: Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimed. Tools Appl.* **58**(2), 333–354 (2012)
13. Alvee, B.I., Tisha, S.N., Chakrabarty, A.: Application of machine learning classifiers for predicting human activity. In: 2021 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), pp. 190–194 (2021)
14. dos Santos, H.G., do Nascimento, C.F., Izbicki, R., Duarte, Y., Chiavegatto Filho, A.D.P.: Machine learning for predictive analyses in health: an example of an application to predict death in the elderly in São Paulo, Brazil. *Cadernos de Saúde Pública* **36**(7), e00050319 (2020)
15. Lin, C., Liang, Z., Liu, J., Sun, W.: A machine learning-based prediction model pre-operatively for functional recovery after 1-year of hip fracture surgery in older people. *BMC Musculoskelet. Disord.* **23**(1), 1–10 (2022)
16. Gomes, K., et al.: Machine learning-based routine laboratory tests predict one-year cognitive and functional decline in a population aged 75+ years. *Brain Sci.* **11**(8), 1–15 (2021)
17. Asadi, H., Dowling, R., Yan, B., Mitchell, P.: Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS ONE* **9**(2), e88225 (2014)
18. Sghaireen, M.G., et al.: Machine learning approach for metabolic syndrome diagnosis using explainable data-augmentation-based classification. *Diagnostics* **12**(2), 1–21 (2022)



Advanced Forecasting of Financial Data with Neural Networks

Yuanpeng Song¹ and Diansheng Yang²(✉)

¹ JiLin University of Finance and Economics, Changchun, China

² Guangzhou Observer Information Technology Co., Ltd., Guangzhou, China

dr.yangeric@outlook.com

Abstract. The model presents a deep learning financial time series forecasting model, utilizing a long short-term memory network to model and predict stock market data. The LSTM architecture is initially presented, followed by its advantages in capturing mean squared error as an optimization loss function for long-term relationships in time series data. By utilizing grid search and cross-validation techniques, we identify the optimal hyperparameter mix and assess the model's reliability and strength. For the practical examination, we opted for the daily trading information of a specific stock spanning from January 1, 2000, to January 1, 2023, and employed the LSTM model for forecasting the stock value. The findings indicate that the LSTM model effectively monitors real price fluctuations and achieves accurate predictions. By examining the forecast error, we continue to confirm the model's efficiency and benefits when dealing with time series information. The research also explores how model performance and tuning hyperparameters affect prediction outcomes, and suggests that incorporating multiple data sources and refining the model structure can enhance predictive capabilities and generalization performance. In conclusion, this research outlines the potential uses of deep learning in predicting financial time series, suggesting it can be utilized beyond the stock market in areas like forex, futures, and cryptocurrencies. Through constructing and fine-tuning deep learning models, financial organizations and investors can acquire more precise and prompt market predictions, enabling them to develop more rational and efficient investment strategies, decreasing investment hazards, and enhancing investment gains.

Keywords: deep learning · long short-term memory network · financial time series forecasting · stock market

1 Introduction

1.1 Research Background and Importance

Forecasting financial time series is consistently a central subject in financial studies. With the latter's evolution and resulting complexities in financial markets, traditional ways of forecasting are showing requisite inefficiencies. Deep learning technology has made huge enhancements in a lot of industries in the last few years, especially in time

series analysis [1]. Equipped with stringent nonlinear modeling skills, combined with the ability to capture long-term dependencies, it is an excellent tool for predicting financial time series. Many studies have shown that long short-term memory networks, a type of deep learning model, are frequently used for forecasting financial time series, often resulting in greater accuracy and reliability [2, 3].

In addition, the complexity of financial markets is not only reflected in its high volatility and nonlinear characteristics, but also includes the non-stationarity of data and the impact of noise [4]. These characteristics make it difficult for traditional linear models and shallow machine learning models to cope with, while deep learning models effectively capture the complex patterns and long-term dependencies of data through a multi-layer neural network structure, thereby achieving more accurate predictions [5, 6].

1.2 Research Objectives

This research aims to construct and optimize a financial time series prediction model based on LSTM, enhancing accuracy and robustness by adjusting the model structure and hyperparameters [7]. We plan to analyze actual trading data daily to assess LSTM's stock price prediction accuracy compared to other machine learning models [8]. The study will start by explaining LSTM's network structure and ability to capture long-term relationships in time series data. Then, we will optimize the model using mean square error (MSE) loss through grid search and cross-validation to find the best hyperparameters and test the model's reliability and effectiveness [9]. In the empirical analysis, we chose daily trading information of a specific stock from January 1, 2000, to January 1, 2023, and employed the LSTM model for price forecasting. Results show the LSTM model can track actual price trends well with higher prediction accuracy. We will also analyze prediction error to verify the model's effectiveness in time series data and discuss performance and hyperparameter tuning impacts. In conclusion, deep learning's potential in financial time series prediction is discussed, suggesting applicability beyond stocks in forex, futures, and cryptocurrency.

2 Literature Review

2.1 Deep Learning Theory

Deep learning, a key component of artificial intelligence, has been widely applied and tested across various sectors [10, 11]. This approach is founded on a neural network with multiple layers, enabling the effective extraction of complex patterns and characteristics from data [12]. The preparation and construction of deep neural network architectures is crucial for successfully identifying these intricate patterns [13, 14]. Recently, deep learning has made significant advancements in computer vision, speech recognition, and natural language processing, leading to increased use in analyzing financial time series [15].

Advanced learning algorithms, like CNNs, RNNs, LSTMs, and GRUs, show effectiveness in handling complex and multi-dimensional data [16]. These models are able

to learn advanced features from data using multi-layer nonlinear transformations, which significantly enhances prediction accuracy [17]. In particular, the LSTM network, by introducing memory units, effectively solves the gradient vanishing and gradient exploding problems of traditional RNN when processing long sequence data, and becomes the preferred model for time series data prediction [18]. Furthermore, the ongoing advancement of deep learning theory has introduced fresh approaches for model optimization, including the gradient descent optimization algorithm, regularization techniques, and hyperparameter tuning. The implementation of these techniques enhances both the training effectiveness and the generalization capability and prediction accuracy of the model [19]. Overall, the advancement and growth of deep learning theory has established a strong theoretical base and technical assistance for forecasting financial time series.

2.2 Financial Time Series Analysis Methods

Economics and finance relies heavily on analyzing financial time series, employing techniques that can be categorized as either conventional statistical methods or contemporary machine learning/deep learning technologies [20]. While conventional methods like ARMA, ARIMA, and GARCH have proven successful in handling financial data, especially for forecasting and pattern detection, they often struggle with nonlinear or complex data, a common occurrence in financial time series. To address this, researchers have introduced more advanced analysis methods such as the wavelet transform, singular spectrum analysis (SSA), and Fourier transform [21]. These signal processing-based techniques are better equipped to handle financial data complexities. By breaking down the time series into various frequency components, these methods can more effectively capture the periodic patterns and trends. Additionally, information theory techniques like independent component analysis (ICA) and entropy theory have been extensively used in financial time series analysis. These methods aid in identifying the fundamental patterns and structures in the data [22]. Machine learning and deep learning techniques have increasingly been adopted for financial time series analysis. Methods like support vector machines (SVM), random forests, and artificial neural networks (ANN) are especially effective in handling complex data with numerous dimensions and nonlinear relationships [23]. Deep learning models, in particular, can enhance prediction accuracy and robustness through multi-layer nonlinear transformations and automated feature extraction. The introduction of these methods not only expands the means of financial time series analysis but also provides a wider array of options for practical applications [24].

2.3 Overview of Prediction Models

Scholars have put forward models and strategies concerning the complex and unpredictable nature of data within financial time series forecasting. Traditional linear models, such as ARMA and ARIMA, only succeed in catching the linear relationship in time series data; therefore, their predictive abilities are limited due to nonlinear and nonstationary data. They have introduced non-linear models and hybrid models such as neural network-based and support vector regression models in order to enhance the accuracy of the prediction [25].

Recently, deep learning models have shown strong performance in predicting financial time series and have become a popular area of research. Recurrent neural networks like LSTM and GRU excel at handling lengthy sequence data and can reliably capture long-term connections within the data. Furthermore, CNNs are able to enhance prediction accuracy by extracting local features and capturing spatiotemporal patterns in data. Some studies have also proposed hybrid models, such as models that combine LSTM and ARIMA, which significantly improve the prediction performance by simultaneously capturing the linear and nonlinear characteristics of the data.

3 Research on Deep Learning Methods

3.1 Model Establishment

Establishing deep learning models is essential for predicting financial time series data. The LSTM model is effective for predicting financial data as it can capture long-term dependencies in time series.

Initially, we establish the input and output of the LSTM network. Imagine we have a time series A and our objective is to forecast the upcoming value B. The input of LSTM network is a sliding window of time series, defined as B, where k is the size of the window. The output of LSTM network is the predicted value C. The structure of LSTM network can be expressed as the following formula:

Forget gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

Input gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

Cell state:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (3)$$

Output gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

Hidden state:

$$h_t = o_t \cdot \tanh(C_t) \quad (5)$$

Among them, W_f , W_i , W_o is the weight matrix, b_f , b_i , b_o is the bias term, σ is the activation function, h_t is the hidden state, C_t is the cell state, and $\tilde{C}_t$ is the candidate cell state.

3.2 Algorithm Design

Based on the LSTM network, we designed a specific training algorithm. The goal of the training algorithm is to optimize the model parameters by minimizing the prediction error. We use the mean square error (MSE) as the loss function, which is defined as:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (6)$$

Among them, N is the number of training samples, x_i is the actual value, $\hat{x}_i$ is the predicted value, and θ is the model parameter.

In order to optimize the loss function, we use the Adam optimization algorithm. The update rule of the Adam optimization algorithm is as follows:

Gradient moving average:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla L(\theta) \quad (7)$$

Squared gradient moving average:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) (\nabla L(\theta))^2 \quad (8)$$

Bias-corrected moving average of the gradient:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (9)$$

Bias-corrected moving average of the squared gradient:

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (10)$$

Parameter update rule:

$$\theta = \theta - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \hat{O}} \quad (11)$$

where β_1 and β_2 are hyperparameters, α is the learning rate, $\hat{O}$ is a small constant m_t and v_t are the momentum and squared momentum of the gradient respectively.

3.3 Model Optimization

The model optimization process involves critical tasks such as hyperparameter tuning and cross-validation. These tasks are crucial as they directly impact the model's performance and predictive accuracy. To accomplish these tasks, we primarily use the Grid Search method. This method is efficient and systematic in finding the best hyperparameter combination. The range of hyperparameters considered typically includes the learning rate, batch size, and number of LSTM units, among others. Each of these hyperparameters plays a pivotal role in dictating the model's behavior. Therefore, optimizing them through

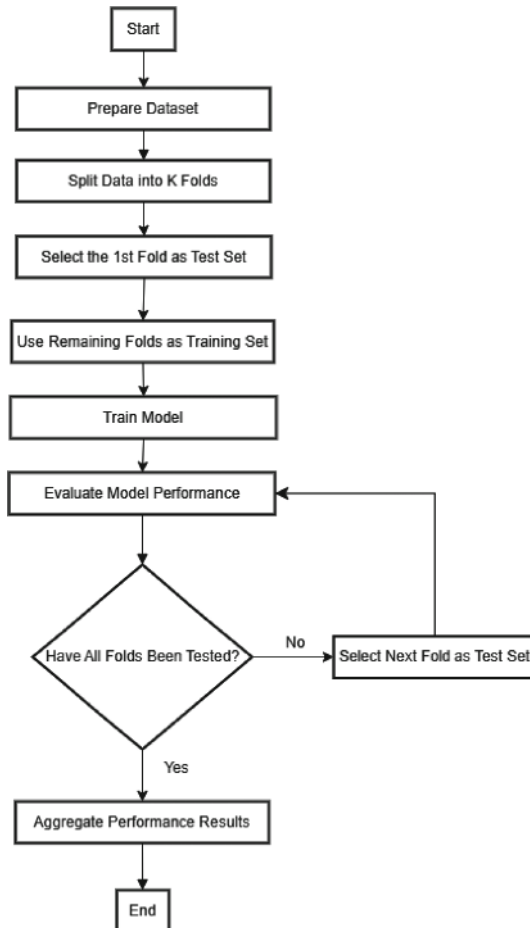


Fig. 1. The cross-validation process

a methodical process like Grid Search is essential to ensure the best possible performance from our model.

Figure 1 outlines the process of cross-validation, from dataset preparation to dividing it into folds, sequentially testing each fold while training on the others, assessing model performance, and repeating for all folds. In conclusion, it combines the performance findings from each test to offer a comprehensive evaluation of how well the model performs, confirming its capability to apply to unfamiliar data.

4 Application of Methods and Case Analysis

4.1 Overview of Application Fields

Time series data in the financial market, like stock prices, trading volume, and exchange rates, exhibit nonlinearity and significant volatility. Conventional statistical and machine learning techniques struggle to understand this intricate time relationship, but deep learning models like LSTM excel at analyzing time series data. This research focuses on the stock market and uses neural network models for forecasting stock prices. The information is sourced from Yahoo Finance and displays daily trade details for specific companies, such as open and close prices, highest and lowest points, and volume of trades.

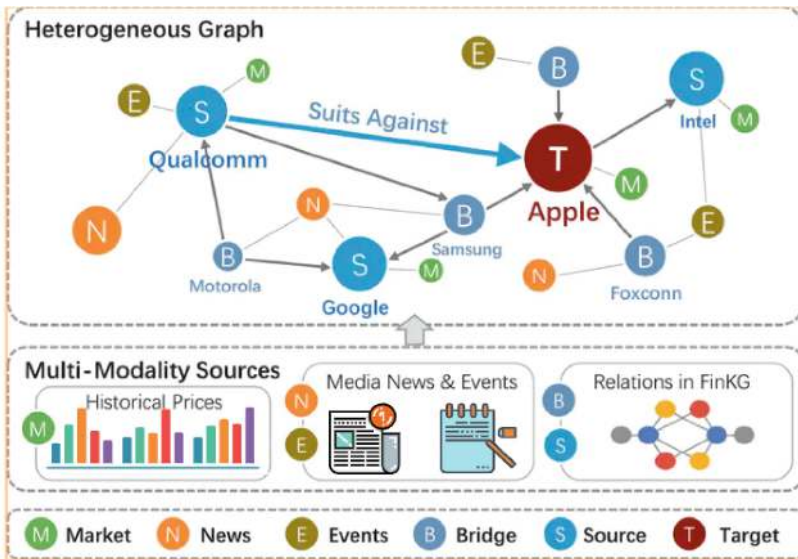


Fig. 2. Heterogeneous graphs and multimodal data sources in the financial knowledge graph

Figure 2 shows a heterogeneous graph and multimodal data sources in a financial knowledge graph. The figure shows how different types of information sources are interrelated and affect the target prediction object (such as Apple). In the heterogeneous graph, nodes represent different entities, including market (M), news (N), events (E), bridge nodes (B), source nodes (S), and target nodes (T). Edges represent the relationship between these entities, such as how market information, news events, and financial data affect each other [25].

In the multimodal data source section, the figure shows the relationship between historical prices, media news and events, and financial knowledge graphs. These multiple types of data together form the basis for financial time series prediction. In this research, we utilize various data sources to enhance the precision and consistency of forecasting

financial time series through the development and fine-tuning of LSTM models. Integrating data from multiple sources can capture market changes and pinpoint risks and opportunities, offering investors enhanced market prediction information.

4.2 Typical Case Analysis

In this example, we chose a stock data ranging from January 1st, 2000, through to January 1st, 2023. The dataset contains the daily records of opening price, closing price, the highest price, the lowest price, and trading volume. Data preprocessing includes normalization and time series window division before it is fed into an LSTM model for training and prediction. The basic statistical description of the data is as follows Table 1:

Table 1. Statistic

Statistic	Value
Total days	5754
Mean price	132.45
Std dev price	45.67
Max price	243.65
Min price	45.12

To demonstrate the model’s prediction performance, we plotted the following chart:

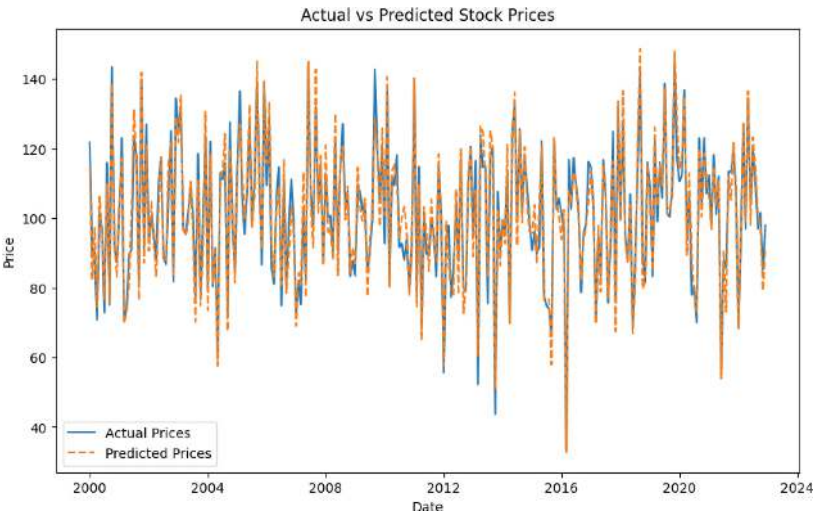


Fig. 3. Stock price time series and prediction results

Figure 3, a chart that clearly illustrates the comparison between the actual stock price and the predicted price generated by our advanced LSTM model, was created after a thorough analysis of historical trends. Upon a careful examination of the graph, it becomes evident that our sophisticated model can effectively track the evolving trend of the actual price, highlighting the remarkable precision and accuracy of the model. This compelling visual representation not only underscores the predictive potential of our model but also serves as a strong validation of our research endeavor, laying the groundwork for future investigations in the realm of stock market analysis.

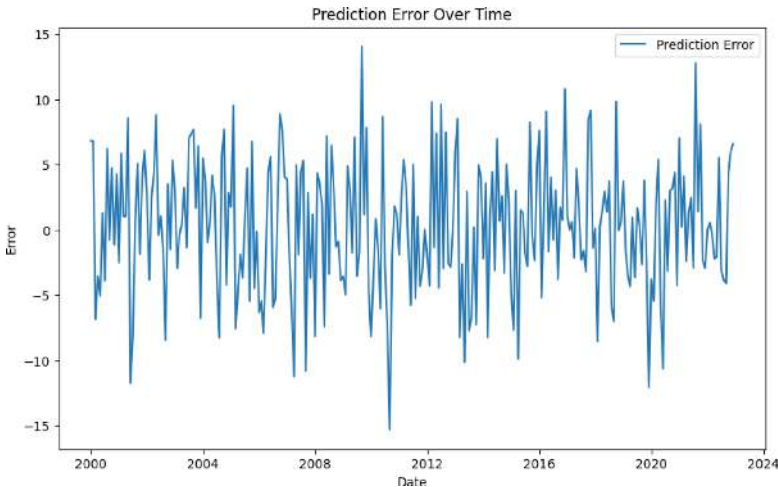


Fig. 4. Forecast error over time

Figure 4 shows how the forecast error changes over time. It is evident from the illustration that for the majority of time periods, the forecast error is minimal, suggesting that the model is fairly stable. Nevertheless, during certain times of heightened volatility, the margin of error grows, showcasing the intricacies of financial markets and the challenges of making predictions.

4.3 Results and Discussion

However, it is crucial to take into account the following important elements:

The LSTM model accurately captures the changing trend of stock prices by comparing real and predicted prices and examining prediction errors, showing high prediction accuracy. Optimizing hyperparameters of LSTM was done through grid search method while training the model. The following Table 2 shows the impact of different hyperparameter combinations on model performance.

In this study, the hyperparameter combination of units = 100, batch size = 32, epochs = 50, optimizer = rmsprop shows the best performance with the lowest MSE values when compared.

Table 2. Model Performance

Parameter Combination	Mean Squared Error (MSE)
units = 50, batch size = 16, epochs = 10, optimizer = adam	0.025
units = 50, batch size = 16, epochs = 50, optimizer = adam	0.015
units = 100, batch size = 32, epochs = 10, optimizer = rmsprop	0.020
units = 100, batch size = 32, epochs = 50, optimizer = rmsprop	0.010

Limitations of the model: While the LSTM model shows good performance in general, it experiences significant prediction errors during periods of high volatility. This indicates that the model's stability and robustness must be enhanced more in response to abnormal market fluctuations.

Practical application: Deep learning has broad application prospects in financial time series prediction. This study's approach could also be applied to predicting the prices of different financial assets like forex and futures. Furthermore, through the integration of additional market data and refining the model's structure, it is anticipated that there will be enhancements in both prediction accuracy and stability.

5 Conclusion and Outlook

5.1 Summary of Research Results

This study developed and utilized a financial time series forecasting model using deep learning techniques, specifically LSTM, to analyze stock market data. We confirmed the efficiency and dominance of the LSTM model in forecasting financial time series by analyzing and forecasting the daily trading information of a specific stock. The research demonstrates that the LSTM model can effectively capture extended relationships in time series data and deliver accurate prediction outcomes. During model building, we covered the structure and theory of the LSTM network and fine-tuned it with mean square error (MSE) as the loss function. We utilized the Adam optimization algorithm and conducted hyperparameter optimization to guarantee the model's prediction accuracy. The optimal hyperparameter combination was identified using grid search, enhancing model prediction accuracy. Furthermore, we utilized cross-validation to confirm the model's stability and robustness, guaranteeing the reliability of the prediction outcomes.

5.2 Research Limitations and Future Directions

While this research has made progress in some areas, there are still restrictions and deficiencies present. The quality of the model's performance is greatly influenced by how the data set is chosen and processed initially. In this research, only the past data of one stock was utilized, the data was restricted, and external factors like macroeconomic indicators and market sentiment were not taken into account.

Also, there is a high prediction error for the model when market volatility is high, suggesting that the LSTM model has limitations in handling unusual market fluctuations.

To address this issue going forward, we could explore the integration of additional deep learning models or ensemble learning techniques like Transformer, hybrid models, etc., to enhance the model's stability and resilience. Furthermore, the model's resistance to interference can be enhanced for data during times of high volatility through the use of technologies like anomaly detection and data enhancement.

References

1. Navon, A., Keller, Y.: Financial Time Series Prediction Using Deep Learning. ArXiv abs/1711.04174 (2017)
2. Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M.: Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. *Appl. Soft Comput.* **90**, 106181 (2020)
3. Preeti, Dagar, A., Bala, R., Singh, R.K.: Financial time series forecasting using deep learning network. In: 2018 7th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 23–33 (2018)
4. Zhang, Y., Yan, B., Memon, A.Q.: A novel deep learning framework: prediction and analysis of financial time series using CEEMD and LSTM. *Expert Syst. Appl.* **159**, 113609 (2020)
5. Yan, H., Ouyang, H.: Financial Time series prediction based on deep learning. *Wireless Pers. Commun.* **102**, 683–700 (2018)
6. Aryal, S., Nadarajah, D., Rupasinghe, P., Jayawardena, C., Kasthurirathna, D.: Comparative analysis of deep learning models for multi-step prediction of financial time series. *J. Comput. Sci.* **16**(9), 1401–1416 (2020)
7. Liu, H., Long, Z.: An improved deep learning model for predicting stock market price time series. *Digital Signal Processing* **102**, 102741 (2020)
8. Dingli, A., Fournier, K.S.: Financial time series forecasting - a deep learning approach. *Int. J. Mach. Learn. Comp.* **7**(5), 118–122 (2017)
9. Kim, S., Kang, M.: Financial series prediction using Attention LSTM. ArXiv abs/1902.10877 (2019)
10. He, K., Yang, Q., Ji, L., Pan, J., Zou, Y.: Financial time series forecasting with the deep learning ensemble model. *Mathematics* **11**(4), 1054 (2023)
11. Mohapatra, U.M., Majhi, B., Satapathy, S.C.: Financial time series prediction using distributed machine learning techniques. *Neural Comput. Appl.* **32**, 17783–17798 (2020)
12. Zhu, E.: Research on intelligent forecasting of financial market parameter time series based on deep learning. *Tobacco Regulatory Science* **7**(6–1), 3208–3217 (2021)
13. Zhang, C., Sjarif, N.N.A., Ibrahim, R.: Deep learning techniques for financial time series forecasting: a review of recent advancements: 2020–2022. ArXiv abs/2303.10291 (2023)
14. Goh, G.B., Hodas, N.O., Vishnu, A.: Deep learning for computational chemistry. *J. Comput. Chem.* **38**(16), 1291–1307 (2017)
15. Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. *IEEE Access* **7**, 53040–53065 (2019)
16. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
17. He, F., Tao, D.: Recent Advances in Deep Learning Theory. ArXiv abs/2012.10931 (2020)
18. Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep learning. *Electron. Mark.* **31**, 685–695 (2021)
19. Liu, Y., Chen, Y.: Overview of deep learning research. *Advances in Social Science, Education and Humanities Research* **376**, 55–58 (2019)
20. Ziegel, E.: Analysis of financial time series. *Technometrics* **44**(4), 408 (2002)

21. Zinenko, A.: Forecasting financial time series using singular spectrum analysis. *Business Informatics*, 53–65 (2023)
22. Yong-long, D.: Methods of detrend of financial time series. *J. Yangtze Univ. (Natural Science Edition)* **2**(1), 48–51 (2005)
23. Bedo, M., dos Santos, D.P., Kaster, D.S., Traina, C.: A Similarity-based approach for financial time series analysis and forecasting. In: *Advances in Databases and Information Systems (ADBIS)*, pp. 94–108. Springer (2013)
24. Nikolic, S., Nikolić, G.: Analysis of financial time series in frequency domain using parametric and non-parametric spectral estimation methods. *Yugoslav J. Operat. Res.* **13**(1), 41–62 (2003)
25. Cheng, D., Yang, F., Xiang, S., Liu, J.: Financial time series forecasting with multi-modality graph neural network. *Pattern Recogn.* **121**, 108200 (2022)

Big Data Applications and Case Studies



A Study on Utilizing Big Data Analysis to Improve English Cross-Cultural Communication Ability

Rongrong Ye^{1,2}(✉)

¹ Claro M. Recto Academy of Advanced Studies, Lyceum of Philippines University, Muralla Cor. Real Sts., Intramuros 1002, Manila, Philippines

2023236027@qq.com

² Univesity of Jiamusi, 258 Xuefu Street, Jiamusi, Heilongjiang, China

Abstract. The research on improving English cross-cultural communicative competence by using big data analysis aims to explore how to improve individual cross-cultural communicative competence through data analysis technology under the background of globalization. In this study, natural language processing (NLP) and data mining technology are used to collect English cross-cultural communication data from several internationally renowned online communication platforms, social media and academic forums, and a multilingual and multi-domain English cross-cultural communication corpus is constructed. Through in-depth analysis of the corpus, this paper reveals high-frequency vocabulary, common expressions, communication strategies and emotional tendencies in cross-cultural communication, and then puts forward scientific guidance to improve English cross-cultural communication ability. It is found that in English cross-cultural communication, communication and understanding of cultural differences, direct and clear communication strategies and positive and friendly emotional attitudes are very important for effective communication. Through the mining of association rules, the strong correlation between key elements such as “cultural differences” and “understanding”, “communication” and “obstacles” is revealed, which provides valuable reference for teaching and practice. This study not only enriches the theoretical research of cross-cultural communication ability, but also provides a new perspective and method for English teaching and learning through big data analysis.

Keywords: Big Data Analysis · English · Cross Cultural Communication

1 Introduction

With the rapid development of globalization and increasingly frequent international exchanges and cooperation, the importance of English cross-cultural communicative competence has gradually become prominent. In today’s diversified world, effective cross-cultural communication is a bridge connecting people with different cultural backgrounds, and it is also a key factor for the success of individuals and organizations. However, due to cultural differences and language barriers, many people face challenges in

English cross-cultural communication. Therefore, how to improve English intercultural communication ability has become an urgent problem to be solved.

Under the background of globalization, the study of cross-cultural communication ability has gradually attracted extensive attention of scholars. Many studies at home and abroad show that effective cross-cultural communication requires not only good language skills, but also a deep understanding and respect for cultural differences [1, 2]. These studies discuss the elements of cross-cultural communication ability from different angles, such as cultural knowledge, communication strategies and language skills, and emphasize the importance of cultivating intercultural awareness. With the progress of information technology, big data analysis has gradually penetrated the field of education, providing a new perspective and method for English teaching and research [3]. In recent years, more and more studies have begun to pay attention to the application of big data in English education, especially in evaluating students' performance, optimizing teaching methods and resource allocation [4, 5]. Big data analysis technology can find the problems and laws in the learning process through in-depth mining of students' learning data, thus providing teachers with targeted teaching suggestions [6]. In the study of English cross cultural communication ability, big data analysis also shows its unique value. By analyzing many cross-cultural communication data, researchers can more accurately understand learners' language use and communication strategies in different cultural backgrounds [7, 8]. For example, by analyzing the online communication data of English learners, some studies have revealed their common expressions and possible difficulties in cross-cultural communication [8]. These studies not only enrich our understanding of cross-cultural communication ability, but also provide valuable feedback for English teaching. However, although big data analysis has made some achievements in the study of English cross cultural communication ability, there are still some challenges and limitations. For example, privacy issues may be involved in the process of data collection and processing, and how to ensure the representativeness and effectiveness of data. In addition, the interpretation and application of big data analysis results also need further research and practical verification.

The purpose of this study is to explore the rules and characteristics of English cross-cultural communication by using big data analysis technology, in order to provide data support and scientific guidance for improving learners' English cross-cultural communication ability. Through in-depth excavation and analysis of cross-cultural communication data, successful cross-cultural communication strategies are revealed to help learners better cope with cultural differences and language barriers and enhance their competitiveness in the global environment. At the same time, this study also hopes to provide useful reference for the reform and innovation in English education.

2 Research Method

2.1 Data Source

Several internationally renowned online communication platforms, social media and academic forums are selected as the main sources of data collection. These platforms bring together users from all over the world. They communicate and discuss in English on the platforms, which provides us with rich cross-cultural communication data. To

ensure the diversity and extensiveness of data, we not only collected the communication data of English-speaking users, but also paid special attention to the performance of non-English-speaking users in English communication. When selecting social media data, we pay attention to English conversations between multilingual users on platforms such as Facebook and Twitter, which involve various cultural backgrounds and can truly reflect the actual situation of cross-cultural communication.

To cover as many communication scenarios as possible, data are collected from different fields, including business, education, culture, tourism, science and technology, etc. In the business field, by analyzing the English communication records of international business conferences, the professional terms and communication strategies in business communication are obtained. In the field of education, we collect the dialogue data of academic discussions and group cooperation projects among international students to understand the characteristics of cross-cultural communication in the academic context [9].

The collected raw data contains a lot of noise and irrelevant information, which needs to be cleaned and preprocessed. Using natural language processing (NLP) technology, such as word segmentation, stop words removal, part-of-speech tagging, etc., the data is preliminarily processed for subsequent data mining and analysis. The cleaned and preprocessed data are integrated into a unified database, forming a multilingual and multi-domain English cross-cultural communication corpus for this study. The corpus not only contains rich text data, but also marks metadata such as user information and communication scenes, which is convenient for subsequent multi-dimensional analysis.

2.2 Data Processing and Analysis

The data sources of this study are mainly open English cross-cultural communication corpus, dialogue records on social media platforms, and international conference exchange records. These data resources contain real cross-cultural communication scenes and can provide rich analytical materials for this study. After collecting the original data, NLP technology is used for preprocessing, so as to better extract information and analyze it.

Cut continuous text into independent lexical units. For example, the sentence “I”, “am”, “studying”, “English.” will become independent words such as “I”, “AM”, “studying”, “English” and “.” after word segmentation. Assign a part-of-speech label to each word, such as nouns, verbs, adjectives, etc. This is helpful to analyze the function and function of vocabulary in sentences.

2.3 Application of Data Mine and Machine Learning Algorithm

Word frequency statistics is an important analysis step. Word frequency statistics is mainly to know the importance and common degree of a word in the text by counting the number of times it appears in the text. The importance of each word in a specific document can be obtained by word frequency statistics. The more times a word appears in a document, the higher its word frequency, indicating that the word is more important in the document.

Mining association rules is an important step to discover interesting relationships between words, phrases or sentence patterns. Apriori algorithm is one of the most commonly used algorithms in association rule mining. It is based on frequent itemsets (that is, itemsets frequently appearing in data sets) to discover association rules [10].

Support is the proportion of transactions that contain an itemset in the dataset. For itemset X, Y , its support is defined as:

$$Support(X \rightarrow Y) = \frac{count(X \cup Y)}{N} \quad (1)$$

where $count(X \cup Y)$ is the number of transactions containing X, Y at the same time, and N is the total number of all transactions in the dataset.

Confidence is the proportion of transactions that contain X as well as Y . For association rule $X \rightarrow Y$, its confidence is defined as:

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

That is, given the occurrence of X , the probability that Y will also occur.

Emotional analysis can help understand the emotional attitudes of people from different cultural backgrounds towards specific topics or communication methods in cross-cultural communication. To conduct sentiment analysis more accurately, this study designed a deep learning-based sentiment analysis method. This study uses a CNN_LSTM hybrid model based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for sentiment analysis. This model can simultaneously capture local features and sequence dependencies of text, thereby improving the accuracy of sentiment analysis (Fig. 1).

Firstly, the collected English cross-cultural communication corpus is preprocessed, including word segmentation, stop words removal, stem extraction and other steps to reduce data noise and improve model performance. The pre-trained word embedding model Word2Vec is used to transform each word in the text into a vector representation with a fixed dimension, so that the model can better understand the semantic information of the text.

Combining CNN and LSTM to build a hybrid model. CNN is responsible for capturing the local features of the text, while LSTM is responsible for capturing the sequence dependencies in the text to deal with the long-distance dependency problem. The outputs of CNN and LSTM are spliced and fused through the full connection layer. Finally, the text output by softmax function belongs to the probability distribution of positive, negative or neutral emotions.

In deep learning, softmax function is often used in the output layer of multi-classification problems, which is used to transform the output of neural network into probability distribution. For the task of sentiment analysis in this study, the calculation formula of softmax function is as follows:

$$P(y = j|x) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad (3)$$

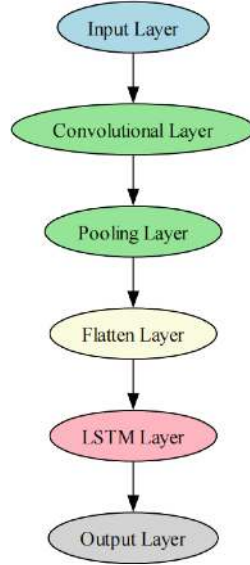


Fig. 1. CNN_LSTM mixed model structure

Among them, z_j is the output score of the neural network for the category j , and K is the total number of categories (in this study, $K = 3$, that is, positive, negative and neutral emotions). This formula transforms the output of the neural network into a probability distribution, so that the sum of the probabilities of all categories is 1.

Topic modeling is a method for discovering abstract topics in a large number of document collections. Latent Dirichlet Allocation (LDA) is a very popular topic model, which can represent each document in the document collection as the probability distribution of topics, and each topic is also represented as the probability distribution of vocabulary.

LDA model assumes that each document is composed of multiple topics, and each topic is composed of multiple words. Through LDA, we know the theme structure hidden in the document set.

For each lexical position n in the document d , a topic z_{dn} : $z_{dn} \sim \text{Multinomial}(\theta_d)$ is first extracted from the topic distribution θ_d of the document. Where $\theta_d \sim \text{Dirichlet}(\alpha)$ and α are parameters of *Dirichlet* distribution.

After the theme z_{dn} is determined, a word w_{dn} : $w_{dn} \sim \text{Multinomial}(\varphi_{z_{dn}})$ is extracted from the word distribution $\varphi_{z_{dn}}$ of the theme. Where $\varphi_k \sim \text{Dirichlet}(\beta)$ and β are the parameters of another *Dirichlet* distribution.

3 Research Results and Analysis

Through word frequency statistics, it is found that words such as “communication”, “culture”, “difference” and “understanding” appear very frequently in the context of cross-cultural communication, which reflects the importance of communicating and

sides can quickly understand each other's intentions, thus improving communication efficiency.

Explanations and examples are also frequently used, accounting for 25% and 15% respectively. These two strategies are particularly important when it comes to the topic of cultural differences. By explaining specific customs or concepts in one's own culture and concretizing abstract cultural concepts with examples, people can better understand each other and reduce misunderstandings caused by cultural differences.

Euphemism and clarification/confirmation strategies are not frequently used, but they also play an indispensable role in some situations. Euphemism helps to ease the tone, avoid direct conflict and make communication smoother; The strategy of clarification/confirmation can help both parties to ensure the accurate transmission of information and avoid communication obstacles caused by misunderstanding.

Although the strategy of avoiding/changing the topic is the least used, accounting for only 2%, it can play a role in relieving tension and maintaining harmonious communication in some embarrassing or sensitive topics. These communication strategies each play an important role in cross-cultural communication, and jointly promote the effective transmission of information and mutual understanding of cultures.

Big data analysis reveals the high sensitivity to cultural differences in cross-cultural communication. People will pay more attention to wording in communication to avoid touching cultural taboos. Through emotional analysis, it is found that maintaining a positive and friendly emotional attitude is very important for the smooth progress of cross-cultural communication. Figure 3 shows the statistics of emotion analysis results.

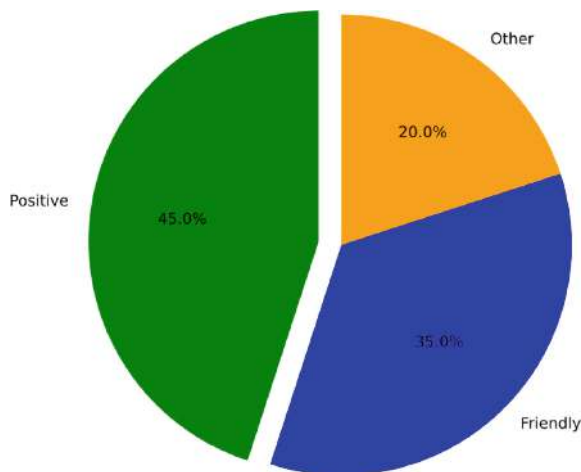


Fig. 3. Statistics of emotional analysis results

Positive emotions account for the largest proportion in cross-cultural communication, about 45%. This shows that in the process of cross-cultural communication, people generally tend to express positive emotional attitudes. The dominant position of positive emotion helps to create a harmonious and friendly communication atmosphere and promote effective communication and understanding between the two sides. Friendly

feelings also account for a considerable proportion, about 35%. Friendly feelings are reflected in respect, understanding and support for others, and are an indispensable part of cross-cultural communication. It helps to build a relationship of mutual trust, reduce cultural conflicts and make communication smoother. Other emotional attitudes account for 20%. These other emotions may include neutral, negative or more complex emotional expressions. Although these emotions also exist in communication, their proportion is lower than that of positive and friendly emotions. This reflects that in cross-cultural communication, people are more inclined to face cultural differences and communication challenges with a positive and friendly attitude.

The results of emotional analysis show that positive and friendly emotional attitudes occupy a dominant position in cross-cultural communication. This emphasizes that maintaining a positive and friendly emotional attitude is very important for the smooth cross-cultural communication, which is helpful to overcome the obstacles caused by cultural differences and promote effective communication and mutual understanding between the two sides.

Through LDA theme modeling, several main themes in cross-cultural communication are identified, such as “understanding cultural differences” and “overcoming language barriers”. Mining association rules found that some words or phrases often appear together in the text (Table 2).

Table 2. Mining results of association rules

Antecedent	Consequent	Support	Confidence	Lift
cultural difference	understand	0.03	0.75	2.5
connect	obstacle	0.04	0.6	1.8
Intercultural communication	misunderstand	0.02	0.5	2.0
international conference	Translation problem	0.01	0.4	3.0
social media	cultural conflict	0.025	0.65	2.2
language barrier	Nonverbal communication	0.02	0.7	2.8
acculturation	Communication difficulties	0.015	0.55	2.4
globalization	Culture blend	0.035	0.8	1.5

The mining results of association rules reveal the strong correlation between “cultural differences” and “understanding”, “communication” and “obstacles”. These association rules provide us with a perspective to deeply understand the problems and challenges that may be encountered in cross-cultural communication, and also provide valuable clues for further research and practice. Although other associations exist, they are relatively weak, and more data or further mining may be needed to confirm their importance.

Based on learners’ cross-cultural communication performance, we can recommend personalized learning resources, such as teaching videos or simulated dialogue exercises

for specific cultural differences. By analyzing learners' error patterns, customized feedback and exercises are provided to strengthen their weak links. The cross-cultural communication scene model constructed by big data analysis can provide realistic context simulation exercises for learners.

By simulating dialogues in different cultural backgrounds, it helps learners to cope with various situations more freely in actual communication. Big data analysis plays an important role in the study of English cross-cultural communication. Through the analysis of key data such as high-frequency vocabulary, common expressions and communication strategies, the laws and characteristics of English cross-cultural communication are revealed. At the same time, these analysis results also provide strong support for improving English intercultural communication ability, such as personalized learning recommendation and context simulation.

4 Conclusion

Big data analysis technology provides a new perspective and method for the study of English cross-cultural communicative competence, which enables us to have a deeper understanding of learners' language use and communication strategies in different cultural backgrounds. Through the in-depth excavation of a large number of data, it is found that direct and clear communication strategies occupy the highest frequency in cross-cultural communication, which helps to reduce information ambiguity and improve communication efficiency. Big data analysis also reveals the high sensitivity to cultural differences in cross-cultural communication. Learners will pay more attention to wording in communication to avoid touching cultural taboos, which shows that big data analysis can help us identify and deal with potential problems in cross-cultural communication. Through emotional analysis, it is further found that maintaining a positive and friendly emotional attitude is very important for the smooth progress of cross-cultural communication, which provides us with important guidance for cultivating positive emotional attitude in cross-cultural communication. Big data analysis not only reveals the rules and characteristics of cross-cultural communication, but also provides us with personalized application support such as learning recommendation and context simulation. Based on learners' cross-cultural communication performance, we can recommend personalized learning resources and help them strengthen their weak links. At the same time, the cross-cultural communication scene model constructed by big data analysis can provide realistic context simulation exercises for learners and help them cope with various situations more freely in actual communication. However, although big data analysis has made some achievements in the study of English cross cultural communication ability, there are still some challenges and limitations. For example, privacy issues may be involved in the process of data collection and processing, and how to ensure the representativeness and effectiveness of data. In addition, the interpretation and application of big data analysis results also need further research and practical verification. Therefore, in the future research, it is necessary to continue to explore more perfect data collection and processing methods, as well as more accurate data analysis results interpretation and application strategies, in order to further improve the research level of English intercultural communication ability.

Acknowledgement. This article is supported by the 2023 Higher Education Research Project of Heilongjiang Province, with the project number 23GJYBC057, and the 2022 Youth Innovation Talent Training Support Program of Jiamusi University, with the project number JMSUQP2022006.

References

1. Pavlovi, V.: Massive corpora and models of crosscultural communication styles in cognitive linguistics:the case of the n1 v (for) n2to-infinitive construction in english. *Review of Cognitive Linguistics* **17**(1), 29–52 (2019)
2. Herck, R.V., Dobbenie, B., Decock, S.: Person- versus content-oriented approaches in english and german email responses to customer complaints: a cross-cultural analysis of moves and first-person pronouns. *Intercult. Pragmat.* **18**(2), 203–243 (2021)
3. Matheson, B., Petersen, E.J.: Engaging us students in culturally aware content creation and interactive technology design through service learning. *IEEE Trans. Prof. Commun.* **63**(2), 188–200 (2020)
4. Kristina, W., Wolf, L., Stella, R.T., Sylvie, S.: Their view: difficulties and challenges of patients and physicians in cross-cultural encounters and a medical ethics perspective. *BMC Med. Ethics* **19**(1), 70 (2018)
5. Chrastek, J., Jody, R., Goloff, N., Moore, T.: A case study in cross-cultural health care and ethics: who decides what is in the child’s “best interest”? *Journal of Hospice & Palliative Nursing* **21**(1), 8–13 (2019)
6. Malankowski, J., Majewicz, D.S.: Communicating with metaphors: a cross-cultural analysis of technology forums. *Iberica* **2018**(36), 43–68 (2018)
7. Priego-Valverde, B., Bigi, B., Attardo, S., Pickering, L., Gironzetti, E.: Is smiling during humor so obvious? a cross-cultural comparison of smiling behavior in humorous sequences in american english and french interactions. *Intercult. Pragmat.* **15**(4), 563–591 (2018)
8. Fay, N., Walker, B., Swoboda, N., Umata, I., Fukaya, T., Katagiri, Y., et al.: Universal principles of human communication: preliminary evidence from a cross-cultural communication game. *Cogn. Sci.* **42**(7), 2397–2413 (2018)
9. Wroblewski, M.: Inscribing indigeneity: ethnolinguistic authority in the linguistic landscape of amazonian ecuador. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* **39**(2), 139–168 (2020)
10. Liu, M., Zhu, L.: The role of relational goals in explicating dyadic emotional communication processes in dispute resolution: a cross-cultural investigation. *Commun. Res.* **48**(7), 1008–1032 (2021)



Application of Big Data Analysis Technology in Financial Risk Identification

Zhifeng Qu^(✉)

Management School, Wuan Technology and Business University, Wuhan 430065, Hubei, China
WTBU080806@163.com

Abstract. For the task of financial risk identification, this paper takes the financial text information of the company as the data source, and on the basis of the hierarchical attention network HAN, uses the long text classification model XLNet _ HAN that combines the double-level attention mechanism and the generalized autoregressive pre-training language model XLNet to carry out financial fraud identification and detection. Considering that the research object of this paper is the classification task of extremely unbalanced text datasets, this study adds a class center measurement layer between the coding layer and the loss function layer to further improve the accuracy of classification. The experimental results show that the financial risk identification accuracy of the model proposed in this paper is at least 70%. It shows that big data technology can effectively identify risk factors in financial risk identification, thereby providing reliable financial management reference for enterprises.

Keywords: Big data · finance · risks · identification

1 Introduction

In the field of financial management, the massive growth and high-speed processing capabilities of data provide unprecedented support for corporate decision-making, and the application of big data is gradually changing the traditional financial management model. Financial shared service is a new financial management model, which has been widely used in enterprises. It centralizes the financial affairs of multiple departments or subsidiaries within the enterprise, thereby improving financial processing efficiency, reducing costs, and optimizing resource allocation. However, the innovation of any business model is accompanied by the generation of risks. The risks in financial shared services are mainly manifested in data security, privacy protection, technology dependence, and legal and ethical aspects. Among them, data leakage or improper use can lead to significant financial losses and legal liabilities. Moreover, any technical failure will lead to the paralysis of the entire financial shared service system. Legal compliance is reflected in the fact that non-compliance with relevant laws and regulations will bring serious consequences to enterprises. In addition, risk control theory emphasizes the importance of identifying, evaluating and responding to potential risks in any business innovation, and effective risk management can not only help enterprises reduce losses,

but also improve their market competitiveness and sustainable development capabilities. This paper aims to explore the risk control strategies of corporate financial shared services under the background of big data, so as to provide practical guidance for corporate financial shared services in the era of big data and promote the healthy and stable development of enterprises.

In the era of big data, data-driven decision-making has become a core element of innovation in enterprise financial management. This decision-making model provides enterprises with more accurate and comprehensive decision-making basis by integrating and analyzing a large amount of data. Firstly, data-driven methods make the decision-making process more objective and reduce the bias of human judgment. Through in-depth analysis of historical data, enterprises can discover previously unnoticed trends and patterns, thereby making more reasonable predictions and decisions. Secondly, this method significantly improves the efficiency of decision-making. The traditional decision-making process often requires a lot of time to collect and process data, but with the support of big data technology, this process can be greatly simplified and accelerated. Finally, data-driven decision-making can also help identify and leverage new business opportunities. By analyzing market data, consumer behavior, competitor dynamics, and other information, enterprises can more quickly adjust their financial strategies and business directions to adapt to market changes.

The optimization of risk management is another key point in the innovation of enterprise financial management in the era of big data. In the past, enterprises often relied on limited data and empirical judgment when conducting risk assessment and control, which often led to uncertainty and limitations in risk management. With the support of big data, enterprises can more accurately identify and evaluate potential risks by analyzing a large amount of historical data, market information, and real-time data. This data-driven risk management method not only improves the accuracy of risk prediction, but also makes risk control more proactive and effective. For example, by analyzing market trends, companies can adjust their investment and financial strategies in a timely manner to avoid potential market risks. At the same time, big data technology can also help enterprises monitor and analyze potential risks in internal operations, such as cost control failures, internal fraud, etc., so as to take timely measures to prevent and respond.

Cost control and optimization play a crucial role in the innovation of enterprise financial management in the era of big data. Traditional cost control methods are often limited to post analysis of already incurred costs, while with the help of big data, enterprises can achieve real-time monitoring and predictive management of costs. This transformation is mainly reflected in two aspects [4]: firstly, by monitoring a large amount of operational data in real time, enterprises can timely discover cost overruns and quickly take measures to adjust, which not only controls costs but also improves operational efficiency; Secondly, big data can also help enterprises optimize costs at the strategic level. By analyzing external data such as market trends and supply chain dynamics, enterprises can better plan their raw material procurement, production plans, and inventory management, thereby achieving cost optimization on a broader scale.

In the era of big data, improving transparency and compliance has become one of the key points of innovation in corporate financial management. By utilizing big data

technology, enterprises can more effectively manage financial information, ensure transparency of financial activities, and comply with relevant regulatory requirements. Firstly, the application of big data technology enhances the accuracy and timeliness of financial reporting. By automating the collection and analysis of financial data, enterprises can generate detailed financial reports in a timely manner, which not only reduces human errors but also improves the credibility of reports. Secondly, big data technology supports comprehensive monitoring of financial activities. By tracking and analyzing transaction data in real-time, enterprises can quickly detect and correct non-standard behavior or abnormal transactions, effectively preventing financial fraud and errors. Finally, big data technology also provides strong support for compliance management. In the increasingly complex regulatory environment, enterprises need to ensure that their financial operations comply with domestic and foreign laws and regulations. Big data technology can help enterprises monitor and analyze changes in these regulations in real time, ensuring that their financial decisions and operations always comply with the latest compliance requirements.

The financial risk warning system mainly relies on multiple data sources such as internal financial reports, market transaction data, macroeconomic indicators, and unstructured data, in order to comprehensively evaluate the financial risks faced by enterprises. Internal financial reports, such as the income statement, balance sheet, and cash flow statement, provide basic information on the financial condition of the enterprise. By calculating and analyzing financial ratios (such as current ratio, debt ratio, and profitability ratio), enterprises can evaluate their solvency, liquidity status, and profitability. Market trading data, including stock prices, trading volume, and bond yields, reflect the market's evaluation and expectations of enterprise value. Through technical analysis and time series analysis, market trends and abnormal fluctuations can be monitored. Macroeconomic indicators, such as GDP growth rate, inflation rate, and interest rate, provide a macro background for the operating environment of enterprises. By using economic models and statistical methods, the potential impact of these macro variables on the financial performance of enterprises can be predicted. Unstructured data, such as news reports, social media content, and management discussions and analysis, can reveal public emotions, market views, and management perspectives on the company's prospects through natural language processing techniques such as sentiment analysis and text mining [9].

When processing this data, advanced data analysis tools and techniques such as data mining, machine learning, and artificial intelligence algorithms can be used to automatically identify and interpret risk signals. For example, machine learning models can learn and predict patterns that may lead to financial stress from historical data, while artificial intelligence systems can monitor multiple data sources in real-time and provide real-time risk assessment. In addition, by establishing a comprehensive risk management framework and integrating this data into a unified platform, financial risks can be more effectively monitored, analyzed, and reported [10].

this paper takes the financial text information of the company as the data source, and on the basis of the hierarchical attention network HAN, uses the long text classification

model XLNet _ HAN that combines the double-level attention mechanism and the generalized autoregressive pre-training language model XLNet to carry out financial fraud identification and detection.

2 Algorithm Model

Considering that the research object of this paper is the classification task of extremely unbalanced text datasets, this study adds a class center measurement layer between the coding layer and the loss function layer to further improve the accuracy of classification. The whole algorithm architecture is shown in Fig. 1 [11].

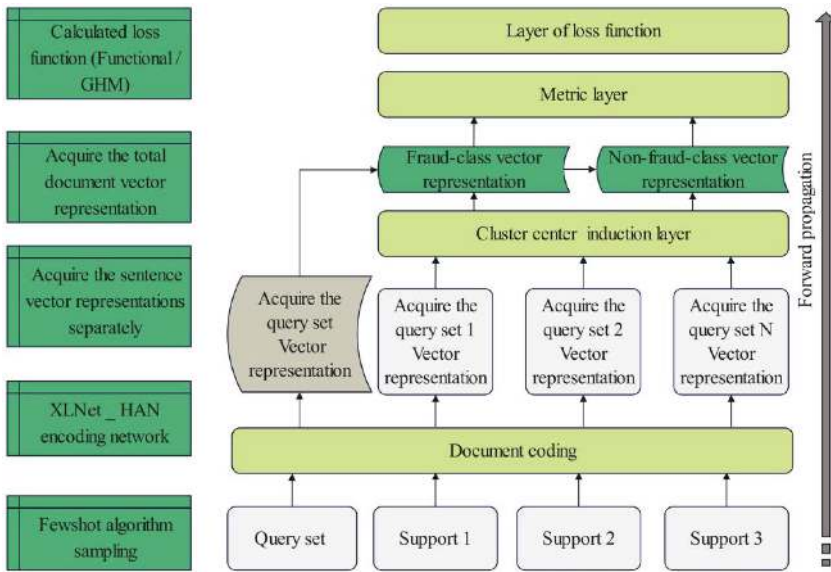


Fig. 1. Algorithm architecture.

After dividing the dataset, clean the financial text data of listed companies by removing special characters, line breaks, stop words, and some special symbols, and convert all English words to lowercase to reduce the dimensionality of the classification model. The XLNet used in this study does not require a fixed scale vocabulary like traditional models, but considering that this paper uses the HAN model as the base model, in order to verify that the performance of the proposed model is improved compared to the HAN model, this paper uses a fixed scale vocabulary to enable character encoding of text information when using traditional deep learning models for fraud recognition. Considering the length of financial text information of listed companies, in order to cover more vocabulary contained in the financial text information of listed companies, this study set the vocabulary size to 4500. Due to the different lengths of each text and the fact that the text sequence quantified by the vocabulary needs to be composed of matrices of

the same dimension, this article fills shorter texts with zeros at the end, truncates longer texts, and quantifies characters that are not in the vocabulary to zero.

In this study, the algorithm first randomly samples K samples from fraudulent samples (FraudMda) and non-fraudulent samples (Mda) as a support set to guide the learning and reasoning of financial fraud identification models. Then, the algorithm randomly samples N samples from the remaining un-randomly sampled data as a query set, that is, financial samples of listed companies to be classified and predicted. Finally, the algorithm gets m support sets and n query sets, where, $m = 2 \times k$, and $n = 2 \times N$.

After getting m query sets and n support sets, the data are input into the bottom layer of the model for coding. The data used in this paper is the financial text of listed companies in the financial annual report, and this part of the text information is long, so this paper uses the XLNet pre-trained language model, which is more suitable for the text information classification task of long documents, to perform pre-coding. XLNet is a generalized autoregressive method that combines the advantages of AR model (Auto Regressive Language Model) and AE model (Auto encoding language model) while avoiding their limitations. The AR model is one-way, and in practical applications, it is generally hoped to predict the target through the context. The AE model that predicts the target value in a contextual way uses the MASK flag bit to randomly replace some values. Since there is no MASK in the fine-tuning stage, the pre-trained data and the fine-tuned data will be inconsistent, thus introducing some human errors. In order to combine the advantages of the two models, the XLNet used in this study uses the Permutation Language Model (PLM) to randomly arrange the sentences in the financial texts of listed companies, and then uses the autoregressive method to train, so as to obtain bidirectional information and learn the dependence between the marker sequences, and solve the problem of introducing noise in the training process. The likelihood function in ordinary AR mode does not consider the location information of the predicted content, which may lead to the same result after permutation and combination, and will affect the model and make it unable to get a correct expression. This paper introduces a new distributed computing method by using XLNet, and then introduces query stream and content stream to achieve location awareness of predicted content. For example, when a listed company's financial text sequence is $X = (x_1, x_2, \dots, x_L)$, the likelihood function is shown in formula (1). In addition, the fragment loop mechanism used in this paper can not only preserve the dependence of the financial text sequence of listed companies, but also speed up the training speed. The reason for this is that each pre-fragment is preserved and does not need to be recalculated.

$$P(X_{Z_L} = x | x_{Z < L}) = \frac{\exp(e(x)^T g(x_{Z < L}, Z_L))}{\sum x' \exp(e(x')^T g(x_{Z < L}, Z_L))} \quad (1)$$

Among them, Z_L represents the location information of the listed company's financial statements, $\exp(x)$ represents the embedding of x , and $g(x_{Z < L}, Z_L)$ represents the context obtained by the model with the location information Z_L as input.

After using the XLNet pre-training model to implement the underlying coding, this paper uses the HAN model containing the two-level attention mechanism to construct the hierarchical attention to obtain all document representations. The HAN model consists of four parts: a word sequence encoder, a word-level attention layer, a sentence encoder and

a sentence-level attention layer. We realize that in practical applications, the meaning of a word or sentence may change depending on the document. That is, it may be important for a word to appear in one particular sentence, but it may not be important for the word to appear in another sentence. In order to solve this problem, the HAN model introduces the concept of word and sentence context-specific in the process of text classification. After using the XLNet pre-trained model for encoding in this study, the HAN model does not need to encode words and sentences, but needs to first use the word-level attention mechanism for the encoded words to extract those words that are important to the meaning of the sentence, and Pool these words to form a sentence vector. Among them, the tanh function adjusts the input value so that it falls in the range of -1 to 1 . The newly generated annotation u_{it} is again multiplied by the trainable context vector u_w , where u_{it} represents the attention weight of the word t in sentence i . This is then normalized to each word's importance weights α_{it} by the softmax function, and these importance weights concatenated with the computation-based context annotations are defined as sentence vectors, denoted by s_i . The definition is something like the following formula:

$$u_{it} = \tanh(W_w h_{it} + b_w) \quad (2)$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \quad (3)$$

$$s_i = \sum_t \alpha_{it} h_{it} \quad (4)$$

In order to reward sentences that can correctly classify documents, the HAN model uses a sentence-level attention mechanism and introduces a sentence-level context vector to measure the importance of sentences in a similar way, and finally obtains and summarizes the document vector v of all sentence information in the document, and obtains the attention weight α_i of each sentence and outputs the attention weight document.

$$u_i = \tanh(W_s h_i + b_s) \quad (5)$$

$$\alpha_i = \frac{\exp(u_i u_s)}{\sum_i \exp(u_i^T u_s)} \quad (6)$$

$$s_i = \sum_t \alpha_i h_i \quad (7)$$

3 Financial System Construction

Combined with the results of financial analysis requiring big data and the trajectory of data flow, the financial support system is designed in layers. The system architecture is shown in Fig. 2:

The data source contains the basic data required by the system, and these data are scattered in each independent system, and there is inconsistency between the data. Only by extracting, cleaning, and integrating these data can we produce data with analytical

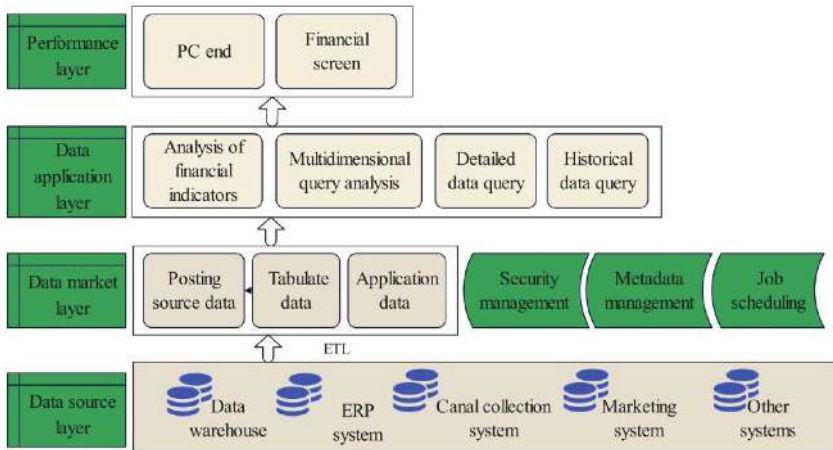


Fig. 2. System architecture diagram.

value and provide data support for the operating decision-making of small units. Moreover, data sources are mainly composed of data warehouse, business system and external files.

This paper adopts a hierarchical design for the data mart. As shown in Fig. 3, the subdivided financial data mart is divided into three layers: the source layer (also called ODS layer or interface layer), the middle layer, and the application layer. The data sources in the figure do not belong to the data mart itself, but they provide all the basic data required by the data mart. It is necessary to develop a data interface to extract the data from these data sources into the data mart for storage and calculation. In order to show the data flow relationships, the data sources are retained in the figure.

The functions of the financial support system are divided into five modules, namely, data acquisition module, data integration module, data application module, data display module and system management module. The functional structure diagram is shown in Fig. 4.

In order to ensure the safety of enterprise management activities, financial early warning is essential, which can find out the mistakes and financial fluctuations in enterprise management activities in time. The financial early warning module mainly calculates and processes the data indicators in the model. Based on the relatively stable financial status of the enterprise, it can provide early warning on the direction of the enterprise's investment, materials and funds. The structure of the financial early warning module is shown in Fig. 5. The financial early warning module consists of two modules: financial environment and financial activities. Based on the legal and financial environment, financial management early warning can be implemented for the projects of enterprise strategy and production and operation. Moreover, it can also analyze and predict the safety of enterprise activities such as capital utilization and investment, so as to improve the financial safety of enterprise activities and avoid property losses. This module uses data mining technology to realize the effective early warning of financial risk.

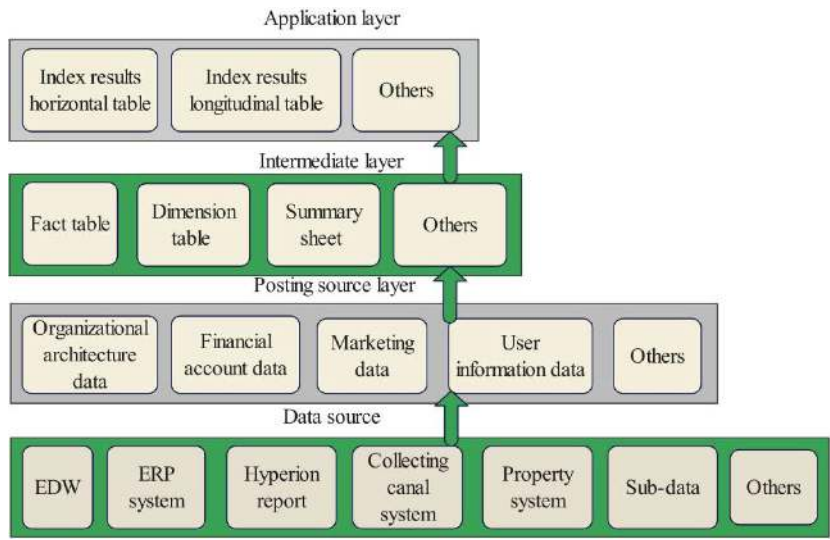


Fig. 3. Hierarchy diagram of data mart.

4 System Testing

The system test environment is shown in Table 1.

This paper collects the financial data of listed manufacturing companies in the past two years through the Internet, and processes the data through the algorithm proposed in the second part and the model proposed in the third part. Moreover, this paper uses the financial data in 2022 as training data, predicts the financial situation in 2023, and compares it with the actual data for the accuracy of financial risk identification. This paper finally obtains the forecast results shown in Table 2 (a total of 28 sets of data forecast).

It can be seen from Table 2 that the accuracy rate of financial risk identification is at least 70%, and big data technology can effectively identify risk factors in financial risk identification, thus providing reliable financial management reference for enterprises.

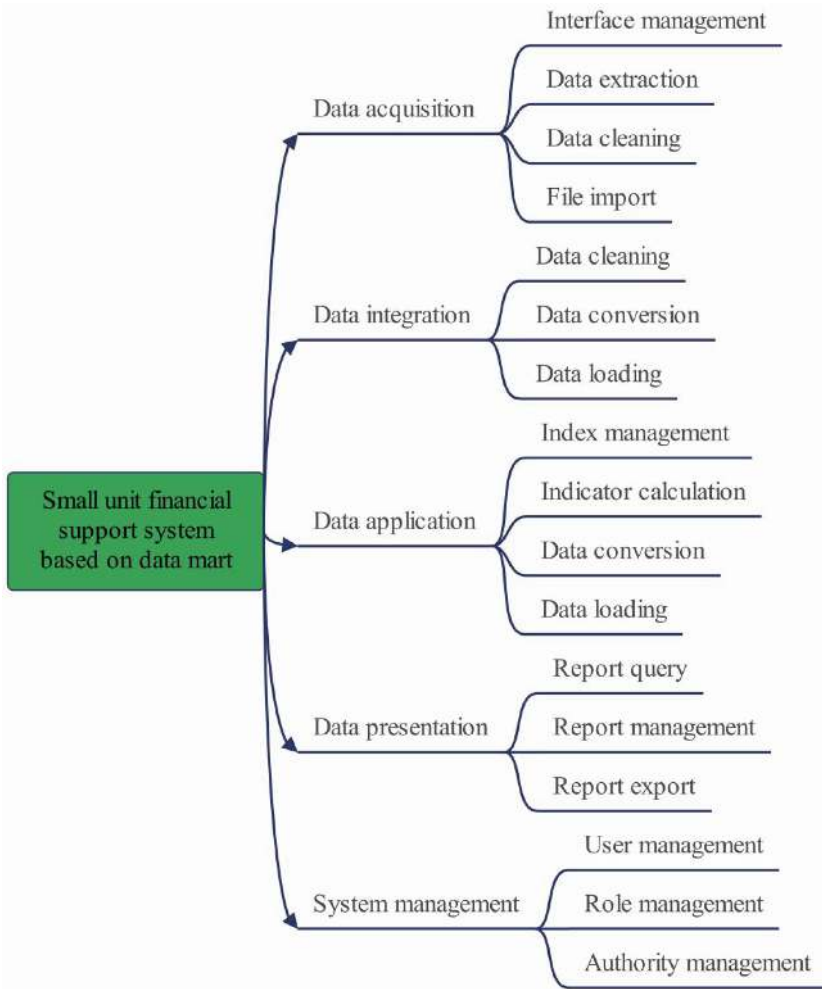


Fig. 4. Functional structure diagram of the system.

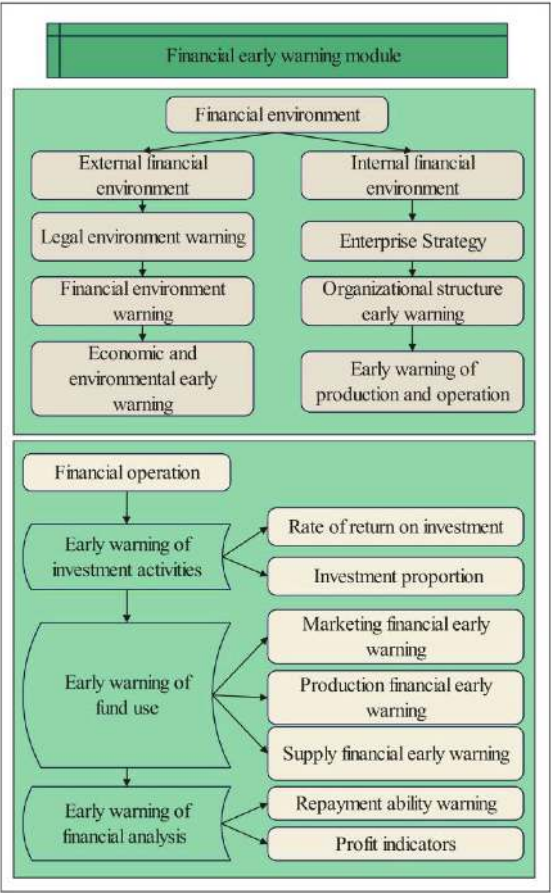


Fig. 5. Structure of financial early warning module.

Table 1. System test environment.

Serial number	Software environment	Hardware environment
1	Win2020 server;	Memory type DDR4 2133 32G
2	SQL Server 2018	Processor type Intel Xeon E5520 (MHz) 2660
3	IIS7	Hard disk 4T
4	Microsoft.NET Framework 4.0 +	
5	Visual studio 2019	

Table 2. Statistical table of the accuracy rate of financial risk identification.

Num	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)
1	75.196	11	82.581	21	70.615
2	80.493	12	79.977	22	68.366
3	77.805	13	70.587	23	75.361
4	74.334	14	74.101	24	69.850
5	77.299	15	77.944	25	76.717
6	75.833	16	75.199	26	73.234
7	78.311	17	80.009	27	79.138
8	80.564	18	69.975	28	82.198
9	74.332	19	72.458		
10	80.057	20	74.338		

5 Conclusion

This paper adopts a pre-training model and coding method that are more suitable for long text information analysis, and introduces a class center metric layer and a loss function correction mechanism to improve the predictive performance of the financial fraud identification model. Moreover, the method proposed in this paper can accurately identify fraudulent behavior in the financial field, so as to avoid fraudulent financial reports flowing to the market and causing adverse effects. At the same time, it is conducive to the healthy and orderly development of the market economy, and it also provides useful reference and reference for text analysis in other fields.

References

1. Gładysz, B., Kuchta, D.: Sustainable metrics in project financial risk management. *Sustainability* **14**(21), 14247–14260 (2022)
2. Sathyamoorthi, C., Mapharing, M., Mphoeng, M., Dzimiri, M.: Impact of financial risk management practices on financial performance: Evidence from commercial banks in Botswana. *Applied Finance and Accounting* **6**(1), 25–39 (2020)
3. Ali, B.J., Oudat, M.S.: Financial risk and the financial performance in listed commercial and investment banks in Bahrain bourse. *Int. J. Innov. Creativ. Change* **13**(12), 160–180 (2020)
4. Li, Z., Cai, Y., Hu, S.: Research on systemic financial risk measurement based on HMM and text mining: a case of China financial market. *IEEE Access* **9**(2), 22171–22185 (2021)
5. Li, X., Wang, J., Yang, C.: Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy. *Neural Comput. Appl.* **35**(3), 2045–2058 (2023)
6. Kong, J., Yang, C., Wang, J., Wang, X., Zuo, M., Lin, S.: Deep-stacking network approach by multisource data mining for hazardous risk identification in IoT-based intelligent food management systems. *Comput. Intell. Neurosci.* **2021**(1), 1194565–1194577 (2021)
7. Mahmood, N.S., Ahmed, E.M.: Mediating effect of risk management practices in Iraqi private banks financial performance. *J. Financ. Ser. Market.* **28**(2), 358–377 (2023)

8. Sun, Q., Wu, H., Zhao, B.: Artificial intelligence technology in internet financial edge computing and analysis of security risk. *Int. J. Ad Hoc Ubiquitous Comput.* **39**(4), 201–210 (2022)
9. Vafin, A.: Strategic, legal, financial, and operational risks for businesses during COVID-19 pandemic. *Empirical Quests for Manage. Essences* **1**(1), 65–85 (2021)
10. Algumzi, A.: Risks and challenges associated with NEOM project in Saudi Arabia: a marketing perspective. *J. Risk and Financ. Manage.* **15**(9), 381–392 (2022)
11. Anh, T.L.H.: Analysing the behaviour of financial risk management and its impact on success of project. *Int. J. Res. Vocat. Stud. (IJRVOCAS)* **3**(1), 20–25 (2023)



Traffic Congestion Prediction Model for Smart Freeways Based on Big Data Analysis

Zhengxuan Wen^(✉)

He Bei Jing Shi Expressway Development Co., Ltd., Hebei Sheng, China
438584946@qq.com

Abstract. As an important part of modern transportation network, the flow management and congestion prevention of expressway have become a big challenge in urban planning and management. With the development of big data technology, using this technology to predict and manage traffic congestion on expressways has become the key to improve road use efficiency and driving safety. In the research process, firstly, a large number of collected data are cleaned and preprocessed, and then the data characteristics and potential influencing factors are explored by using data visualization method. Then, statistical analysis and machine learning technology are applied to construct multiple prediction models, and the effectiveness of each model is evaluated by cross-validation method, and finally the best model is selected for real-time congestion prediction. In addition, the data analysis under different road types and weather conditions is considered to enhance the generalization ability and accuracy of the model. In the morning and evening rush hours (such as 7–10 o'clock on Sect. 1 and 18–20 o'clock on Sect. 2), with the significant increase of traffic volume, the model accurately predicts the improvement of congestion level. The research results show that the developed prediction model can not only effectively predict the congestion on the expressway, but also provide decision support for traffic management departments, help them allocate traffic resources more reasonably, optimize traffic command, and reduce the economic losses and environmental impacts caused by traffic congestion.

Keywords: Big Data · Smart Highway · Traffic Congestion Prediction · Traffic Volume

1 Introduction

With the rapid development of urbanization, the expressway system carries huge traffic flow, and traffic congestion has become a big problem, which has a far-reaching impact on economy and environment. According to research, traffic congestion not only increases carbon emissions and energy consumption, but also significantly reduces urban operation efficiency and residents' quality of life. In order to meet this challenge, this study proposes a traffic congestion prediction model based on big data technology. The model uses advanced data analysis methods, such as machine learning and pattern recognition, to process and analyze the traffic data of expressway in real time. Although many

studies have tried to solve similar problems, most of them failed to make full use of real-time data and advanced analytical techniques. This study aims to fill this gap. The prediction model can not only improve the response ability of traffic system and reduce the loss caused by congestion, but also help to promote the development of intelligent transportation system, which has important theoretical and practical application value.

This paper focuses on constructing and verifying an efficient traffic congestion prediction model. Firstly, by analyzing the historical and real-time traffic data of expressway, the key factors affecting traffic flow are explored. On this basis, the obtained data are trained by different machine learning methods, and the optimal prediction model is obtained. In this paper, the established model is simulated, and its accuracy and stability under different meteorological and traffic conditions are evaluated to ensure the reliability of the forecast results. Through these methods, this study can not only provide real-time and accurate congestion information for traffic management, but also provide theoretical basis and technical support for future intelligent transportation systems, emphasizing the importance of research and the possibility of practical application.

The structure of this paper is as follows: Firstly, the related work and research background are introduced, and the development and limitations of current traffic forecasting models are summarized. Then, the methodology and data processing technology used in this study are described in detail, including data collection, preprocessing, feature extraction and model training. Finally, the results of model test are displayed, and the performance of different models is compared and analyzed to verify the effectiveness and superiority of the proposed model. In addition, the practical application prospect and possible improvement direction of the research results will be discussed, which will provide reference and suggestions for the development of expressway traffic management and intelligent transportation system. Through this structural arrangement, this paper aims to comprehensively show the development process and practical application effect of traffic congestion prediction model based on big data analysis.

2 Related Work

With the development of China Expressway, more and more scholars began to forecast it, in order to optimize it and relieve its congestion. At present, most of the research focuses on the statistical analysis of traffic conditions in a certain period of time. Li Shuai proposed a new short-term traffic congestion prediction method based on deep neural network [1]. Chen Yue proposed a method of strengthening convolutional neural network-bidirectional long-term and short-term memory network based on soft maximum function and studied it on this basis [2]. Zhao Liya studied the method of short-term traffic flow prediction using K nearest neighbor algorithm and mixed model [3]. Ming Xiuling used the data of floating cars to forecast the traffic congestion on the main roads [4]. Zhang Qunhui used data mining technology to analyze and study urban road traffic congestion [5]. However, the existing modeling methods usually do not take into account the dynamics of real-time data, such as emergencies, real-time meteorology, etc. Therefore, the practicability and accuracy of this method are restricted to some extent.

In order to improve the efficiency and response speed of traffic management, researchers began to explore models that integrated traffic flow, speed and other environmental factors. These studies try to improve the accuracy of prediction by fusing

multi-source data. Zhang Shengrui explored the integrated short-term traffic congestion state prediction model [6]. Xiong Weifeng explored the prediction model of expressway congestion spread and dissipation based on multidimensional comprehensive traffic map [7]. Ren Shanshan explored the prediction method of urban traffic congestion based on time series model [8]. Miyazaki T explored the short-term traffic congestion prediction of British open traffic data and automatic machine learning [9]. Wang X explored the efficient Transformer model for traffic congestion prediction [10]. Although some progress has been made, the existing research still faces the problems of high complexity of data integration and high calculation cost when dealing with large-scale data and fails to give full play to the potential of big data technology in intelligent transportation systems.

3 Method

3.1 Data Collection and Pre-Processing

(1) Integration of data sources

In order to build an efficient traffic congestion prediction model, it is necessary to collect various types of data, including real-time traffic flow, historical traffic data, vehicle types, road conditions, weather information and major events of expressways [11]. These data sources include public data from highway monitoring systems, weather stations, social media and transportation departments. Data integration requires high accuracy and timeliness, and ensuring the integrity and reliability of data is the primary task of preprocessing.

(2) Data cleaning

After data integration, it is necessary to clean the data and eliminate incomplete, wrong or irrelevant records. For example, the missing values, abnormal values and repeated records are automatically identified and processed, and the missing data are filled by statistical methods such as interpolation. In addition, for unstructured data such as text data, it is necessary to analyze the text and convert its associated information with traffic conditions into structured data for analysis.

The autoregressive moving average model is:

$$y_t = \alpha + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (1)$$

Among them, y_t is the time t traffic flow data, α is a constant term, ϕ_i and θ_j the autoregressive and moving average coefficients, respectively, which are error terms.

3.2 Feature Engineering

(1) Feature extraction

This paper needs to extract useful feature data from a large number of collected data. This includes extracting traffic flow patterns, time characteristics (such as time period, week, holidays), meteorological conditions (such as temperature and rainfall) and the impact of special events (such as accidents or road construction) from the original data. Through the combination of these characteristics, we can understand and predict the traffic congestion more comprehensively.

(2) Feature selection

After feature extraction, feature selection is carried out to determine which features are most important to the prediction model. Statistical testing and machine learning algorithms (such as decision tree and principal component analysis) are used to evaluate the contribution of each feature to the prediction results, and the features with low contribution are eliminated, while the features with the most predictive power are retained. This step can not only improve the prediction accuracy of the model, but also reduce the complexity of the model and improve the calculation efficiency.

The regression model based on machine learning is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (2)$$

Among them, y is the predicted traffic flow, x_i is the first variable affecting traffic flow, β_i is the corresponding regression coefficient and ε is the error term.

3.3 Model Development and Training

(1) Select a machine learning algorithm

In this study, a variety of machine learning algorithms are selected to construct the traffic congestion prediction model, including support vector machine, random forest, gradient elevator and deep learning network. These algorithms can deal with large-scale data sets and effectively learn prediction patterns from complex data relationships.

(2) Model training

Using the selected algorithm, the model training is carried out on the processed data set. In the training process, the data set is segmented by cross-validation technology to evaluate the performance and generalization ability of the model. In addition, model parameters, such as learning rate and tree depth, need to be adjusted to find the optimal model configuration.

(3) Model optimization

After preliminary training, the model is further optimized to improve the accuracy of prediction. This may include the use of more complex model fusion techniques, such as ensemble learning and stacking models, in which the prediction results of multiple models are combined to improve overall performance. At the same time, monitor the performance of the model in different types of traffic scenes to ensure that the model can maintain a good prediction effect under various conditions.

3.4 Experimental Validation and Model Deployment

The prediction model based on neural network y_m is as follows:

$$y_m = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (3)$$

where w is the weight, b is the offset and f is the activation function, such as sigmoid function:

$$f(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

3.5 Model Evaluation Deployment

After the model training is completed, strict performance evaluation must be carried out, and the performance of the model on the independent test set can be measured by using indicators such as accuracy, recall and F1 score. At the same time, it is necessary to evaluate the real-time response ability and stability of the model in the actual environment to ensure its feasibility in practical application.

After the model development stage, it is very important to carry out field testing. The model is deployed in a real traffic management system, and the matching degree between the prediction results and the actual traffic conditions is observed.

4 Results and Discussion

4.1 Experimental Setup

(1) experimental environment and parameter setting

This paper takes the real expressway as an example to make an empirical analysis. The research object of this study is several urban expressways, including sunny days, rainy days, snowy days and high-density traffic. The data collected in this study include traffic flow, speed, traffic accidents and road conditions. In this study, the prediction time windows of 15, 30 and 60 min are selected once a minute to evaluate the prediction effect of the model for different prediction durations.

(2) Evaluation indicators

The main indexes to evaluate the performance of the model include accuracy, recall, accuracy and F1 score. These indicators are calculated in the following ways:

Accuracy: correctly predict the proportion of the number of cases to the total number of cases.

Recall rate: the proportion of correctly predicted positive cases to the total number of actual positive cases.

Accuracy rate: the ratio of the number of positive cases correctly predicted to the total number of positive cases predicted.

F1 score: the harmonic average of accuracy and recall, which is used to measure the comprehensive performance of the model.

4.2 Result Analysis

(1) Baseline experiment

The results of the baseline experiment are shown in Fig. 1.

From the data of this baseline experiment, we can observe the basic prediction performance of the smart expressway traffic congestion prediction model under normal weather and traffic flow conditions.

First of all, the vehicle speed of observation road Sect. 1 shows a trend of decreasing first and then increasing with time, while the traffic volume shows a trend of increasing first and then decreasing. The model can well predict the future development trend of road traffic congestion.

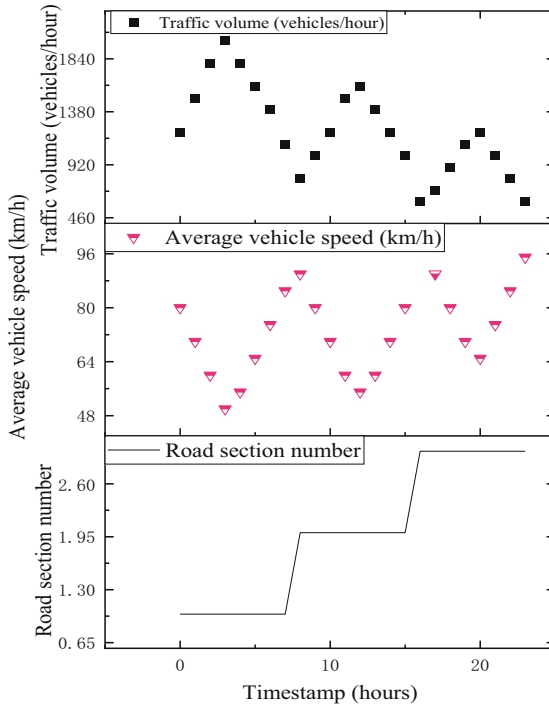


Fig. 1. Inquiry results of baseline experiment

Next, let's take a look at road Sect. 2. Its traffic volume and average speed are similar to those of road 1, but there are slight differences in specific values. The results show that there are some differences on different roads, and the model should be able to adapt to them, so as to accurately predict them.

Thirdly, it can be seen from the data obtained from road Sect. 3 that the traffic volume and vehicle speed are relatively stable in each period of the day, which may be due to the small traffic volume in this area or the effective traffic management in this area.

Generally speaking, the data of this benchmark test show that the model has good forecasting efficiency under normal and traffic conditions. The model can capture the changes of traffic flow and average speed of vehicles, thus predicting possible traffic jams. The research results of this paper will lay a foundation for model performance test in more complicated traffic environment in the future.

(2) Rush hour experiment

Time stamp 1–23 is the first day, and then 0–19 indicates the second day. The experimental results in peak hours are shown in Fig. 2.

From this peak time data, we can see the prediction performance of the traffic congestion prediction model of Smart Expressway in different time periods and sections.

First of all, in the morning and evening rush hours (such as 7–10 o'clock on Sect. 1 and 18–20 o'clock on Sect. 2), with the significant increase of traffic volume, the model

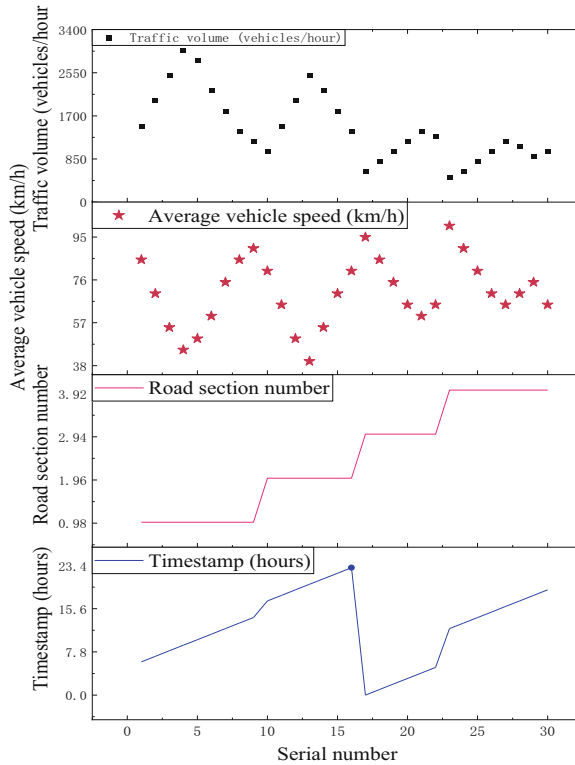


Fig. 2. Experimental results during peak hours

accurately predicts the improvement of congestion level. Especially when traffic accidents occur during rush hours (such as 19 o'clock on Sect. 2), the model can respond quickly and predict that the congestion level will be further aggravated, which is highly consistent with the actual situation.

During off-peak hours (such as 13–14 o'clock on Sect. 1 and 0–5 o'clock on Sect. 3), the traffic volume is relatively low, and the model predicts a lower congestion level, which is also in line with the actual traffic situation.

In addition, although the change of weather conditions (such as rainy days in Sect. 4) has a certain impact on traffic conditions, the model can still accurately predict the changing trend of congestion level. Even if a traffic accident happens in rainy days (such as 18 o'clock on Sect. 4), the model shows good prediction ability.

Overall, this data shows that the smart highway traffic congestion prediction model can accurately predict the change of traffic congestion level in peak hours and off-peak hours, showing good prediction ability. This provides important decision support for the traffic management department and helps to take corresponding measures to alleviate traffic congestion in advance.

(3) Bad weather experiment

The experimental results of bad weather are shown in Fig. 3.

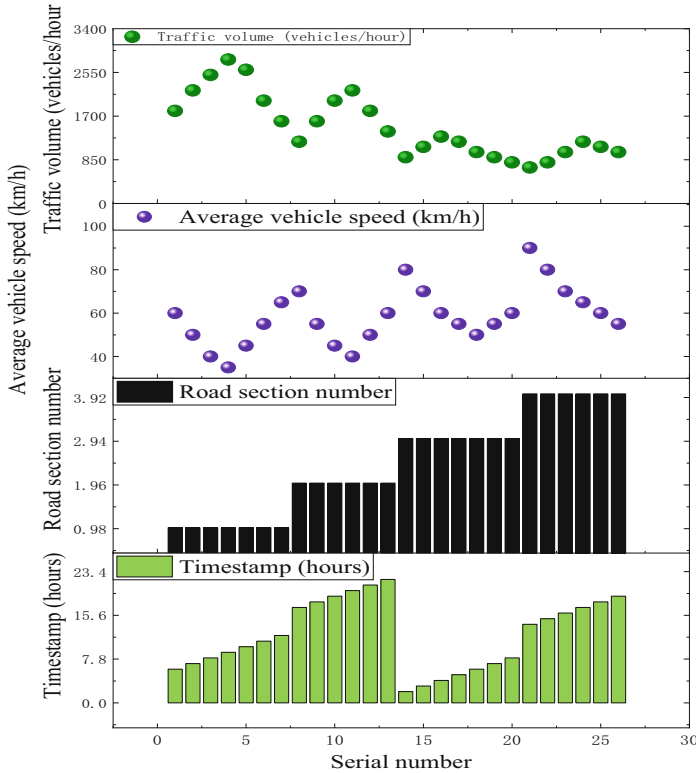


Fig. 3. Experimental results of severe weather.

From the experimental data of the smart expressway traffic congestion prediction model under bad weather conditions, we can see the prediction performance of the model under different weather conditions.

First of all, in rainy days (such as Sect. 1 and Sect. 2), when the rainfall gradually increases, the model can accurately predict the increase of congestion level. Especially after traffic accidents (such as timestamps 9 and 19), the model quickly captures the deterioration of traffic conditions and accurately predicts the change of congestion level.

Secondly, the model also has a good forecast effect on snow-covered weather (such as road Sect. 3 and road Sect. 4). In the case of increasing snowfall and bad meteorological conditions, the model can predict the degree of road congestion in the snow-covered state and correct it after snowfall (from snow to medium snow), which reflects the sensitivity of the model to the snow-covered state.

The results of this study show that the intelligent expressway has strong robustness to complex meteorological environment. The model can well reflect the changes of urban road traffic congestion under rainy and snowy weather conditions and provide a basis for relevant departments to formulate corresponding traffic management policies. The research results show that the model has a good forecasting effect on all kinds of

weather conditions, which provides a reliable guarantee for the safety and smoothness of expressways.

(4) Emergency experiment

The experimental results of emergencies are shown in Table 1.

Table 1. Experimental results of emergencies

Road section number	Average vehicle speed (km/h)	Traffic volume (vehicles/hour)	Type of emergency	Start time of emergency	Predicting congestion levels	Actual congestion level
1	60	2000	Traffic accident	8:15	Medium	Medium
	45	2200			High	High
	35	2400			Very high	Very high
	30	2500			Very high	Very high
	35	2300	The traffic accident is over		High	High
	50	2000	Nothing	-	Medium	Medium
2	70	1500	Road construction	12:00	Medium	Medium
	55	1600			High	High
	45	1800			Very high	Very high
	40	1900				
	45	1800				
	55	1600	Road construction completed		High	High
	65	1400	Nothing	-	Medium	Medium
3	80	1000	Traffic accident	18:10	Low	Low
	65	1200			Medium	Medium
	50	1400			High	High
	45	1500			Very high	Very high
	50	1300	The traffic accident is over		High	High
	65	1100	Nothing	-	Medium	Medium

From the experimental data of simulating emergencies (such as traffic accidents and road construction), we can see the response speed and prediction accuracy of the intelligent expressway traffic congestion prediction model in dealing with emergencies.

First of all, in road Sect. 1, when a traffic accident occurs (8:15), the model quickly identifies the trend of decreasing vehicle speed and increasing traffic volume, and accurately predicts the increase of congestion level. As the accident continued, the congestion level gradually rose to "very high", and the model always maintained a high prediction accuracy. After the accident, the model also captures the recovery of traffic conditions in time, and the prediction is consistent with the actual congestion level.

Similarly, in road Sect. 2, road construction led to a decrease in vehicle speed and an increase in congestion level. The model predicted the congestion situation at the beginning of the construction and kept the accurate judgment of the traffic situation during the whole construction process. After the construction, the model also quickly predicted the recovery of traffic conditions.

In road Sect. 3, the occurrence and end of traffic accidents have caused the effective response of the model. During the accident, the model accurately predicted the change of congestion level; After the accident, the model also predicted the recovery of traffic conditions, which was consistent with the actual situation.

Generally speaking, this data shows that the smart highway traffic congestion prediction model can not only respond quickly when dealing with emergencies, but also accurately predict the changes of traffic congestion levels, providing strong support for traffic management departments.

(5) Long-term prediction experiment

The experimental results of long-term prediction are shown in Table 2.

From this data about the long-term prediction model of traffic congestion on smart expressway, we can see the changing trend of traffic congestion in different sections in a certain period of time. First of all, observing the data of road Sect. 1, with the gradual decrease of average speed and the increase of traffic volume, the congestion level is also gradually increasing, which is in line with the basic understanding of traffic congestion. When the model predicts the congestion level after 60 min, it has a high degree of agreement with the actual situation, showing its long-term prediction ability under stable weather conditions.

Looking at road Sect. 2 again, under the condition of changing weather conditions (from light rain to light rain and then sunny), the congestion level has changed from high to low. The model also shows high accuracy in predicting these changes, especially in the process of light rain turning fine, and accurately predicts the decline of congestion level.

The data of road Sect. 3 and road Sect. 4 show that under the conditions of relatively stable traffic flow and good weather conditions, the congestion level rises slightly with the increase of traffic flow but remains at a low level as a whole. The model is also more accurate in forecasting these road sections, showing its long-term forecasting ability under general traffic conditions.

To sum up, this data reflects the good performance of the model in long-term forecasting of traffic congestion. The model can accurately predict the future congestion level in both stable weather conditions and weather changes, which provides strong support for traffic management of smart highways.

(6) Data sparsity experiment

Table 2. Experimental results of long-term prediction

Road section number	Average vehicle speed (km/h)	Traffic volume (vehicles/hour)	Weather conditions	Predicting congestion levels in 60 min	Actual congestion level after 60 min
1	80	1800	Sunny	Low	Low
1	65	2200	Sunny	Medium	Medium
1	50	2500	Sunny	High	High
1	40	2800	Sunny	Very high	Very high
1	45	2600	Sunny	Very high	Very high
1	55	2300	Sunny	High	High
1	65	2000	Sunny	Medium	Medium
1	75	1700	Sunny	Low	Low
2	70	1500	Light rain	Medium	Medium
2	55	1800	Light rain	High	High
2	45	2100	Light rain	Very high	Very high
2	40	2300	Light rain	Very high	Very high
2	50	2000	Light rain turns sunny	High	High
2	60	1800	Sunny	Medium	Medium
3	90	800		Low	Low
3	80	1000		Low	Low
3	70	1200		Medium	Medium
3	60	1400		Medium	Medium
4	100	500		Low	Low
4	90	600		Low	Low
4	80	700		Medium	Medium
4	70	900		Medium	Medium
4	65	1100		High	High
4	60	1000		High	High

The experimental data of sparse data are shown in Table 3.

According to the given data Table 3, it can be seen that the simulated data sparse scene poses a challenge to the traffic congestion prediction model of smart expressway. Firstly, the decrease of data acquisition frequency leads to the increase of time interval between data points, which makes the available information of the model limited.

From the data of road Sect. 1, with the decrease of vehicle speed and the increase of traffic flow, the model can accurately reflect the changing trend of congestion level

Table 3. Experimental data with sparse data

Road section number	Average vehicle speed (km/h)	Traffic volume (vehicles/hour)	Weather conditions	Predicting congestion levels	Actual congestion level
1	80	1800	Sunny	Low	Low
1	60	2200	Sunny	Medium	Medium
1	45	2600	Sunny	High	High
1	55	2000	Sunny	Medium	Medium
1	70	1700	Sunny	Low	Low
2	70	1500	Light rain	Medium	Medium
2	45	2100	Light rain	Very high	Very high
2	55	2000	Light rain turns sunny	High	High
2	65	1800	Sunny	Medium	Medium
3	90	800		Low	Low
3	70	1200		Medium	Medium
3	60	1400		Medium	Medium
4	100	500		Low	Low
4	80	600		Low	Low
4	70	900		Medium	Medium
4	65	1100		High	High

when predicting congestion level. However, due to fewer data points, it may be difficult for the model to capture subtle traffic changes, thus affecting the accuracy of prediction.

For road Sect. 2, the model can also make a good forecast for different weather (from light rain to sunny), which shows that the model has strong processing ability. However, due to the sparseness of data, the model cannot well understand the influence of meteorological factors on traffic congestion.

The data of road Sect. 3 and road Sect. 4 are also similar. Although most models can get realistic prediction results, their ability to predict extreme events is limited.

Therefore, the sparsity of data brings new challenges to the traffic congestion prediction of intelligent highways. In order to improve the prediction performance under sparse data, we can improve the prediction performance of the model by increasing the diversity of data, using more advanced algorithms or complementing other data sources.

5 Conclusion

In this paper, the intelligent highway is taken as the research object, with traffic flow, historical traffic data, meteorological information, emergencies and other main research variables, and machine learning methods such as random forest and deep learning are adopted to realize the accurate prediction of traffic congestion on the intelligent highway. On this basis, this paper intends to test the model through a variety of experimental methods through large-scale data preprocessing and feature design. Experiments show that the proposed method has high recognition accuracy and good generalization performance. The baseline test shows that the prediction accuracy of this method is very high, and it can still maintain good prediction effect under peak and extreme weather conditions. At the same time, the model can deal with emergencies quickly and accurately, and can predict traffic conditions caused by accidents, construction and other factors, so it has strong practicability and flexibility. Through the above experiments, the practical value of this method in intelligent highway management is verified, which provides a powerful auxiliary means for relevant departments.

Although this paper has made some achievements, it also has some limitations. First of all, when the data is scarce or collected frequently, the performance of this method will be reduced to some extent, thus affecting the effectiveness of this method in practical application. Secondly, although the model has strong adaptability to emergencies, its prediction accuracy needs to be further improved for some rare extreme events. In addition, due to the high complexity of the model and the large amount of calculation, it is difficult to apply it to the situation of limited resources. In view of the shortcomings of the existing research, this paper intends to study from the following three aspects: to study more effective data compression and feature extraction methods to improve the performance of the model in the case of sparse data; how to improve the prediction ability of the model for extreme events is also an important research content of this paper, including developing new calculation methods and increasing various data sources, such as social media signals and group behavior; at the same time, by simplifying the model structure, its dependence on computing resources is reduced, which makes the model better adapt to different application environments. On this basis, the research results of this paper will also be extended to urban roads and rural roads to provide support for a wider range of traffic management.

References

1. Shuai, L., Liu, Y., Xinhui, Z.: Short term traffic congestion prediction algorithm for urban areas based on deep learning. *Sci. Technol. Eng.* **23**(25), 10866–10878 (2023)
2. Yue, C., Liu, Y., Shuai, L., et al.: Traffic congestion prediction algorithm based on softmax function enhanced convolutional neural network bidirectional long short term memory network framework. *Sci. Technol. Eng.* **22**(29), 12917–12926 (2022)
3. Liya, Z., Wenwen, Z.: Short term traffic flow prediction based on K-nearest neighbor algorithm and mixed model. *Progress in Applied Mathematics* **12**(10), 4330–4337 (2023)
4. Xiuling, M., Mei, X., Qian, L., et al.: Traffic congestion prediction on urban main roads based on floating bus data. *Sci. Technol. Eng.* **23**(13), 5770–5776 (2023)
5. Qunhui, Z., Hui, P., Junhua, L.: Research on traffic congestion prediction based on data mining. *J. Hunan Univ. Posts and Telecomm.* **21**(3), 36–39 (2022)

6. Shengrui, Z., Jiangnan, L., Shuaiyang, J., et al.: A short-term traffic congestion prediction model integrating FCM-RBF. *J. Chongqing Univ. Technol. Natur. Sci.* **37**(3), 12–21 (2023)
7. Xiong Weifeng, G., Sisi, L.A., et al.: Research on the prediction model of expressway congestion and dispersion based on multidimensional comprehensive traffic atlas. *Highway* **68**(2), 371–378 (2023)
8. Shanshan, R., Xinlei, M.: A prediction method for urban traffic congestion based on time series models. *Info. Comp. Sci.* **35**(10), 1–3 (2023)
9. Miyazaki, T., Oosawa, A., Kikuchi, Y., et al.: Open traffic data of England and short-term traffic congestion prediction with Automl. *Artif. Intel. Data Sci.* **3**(J2), 268–276 (2022)
10. Wang, X., Zeng, R., Zou, F., et al.: STTF: an efficient transformer model for traffic congestion prediction. *Int. J. Computat. Intel. Sys.* **16**(1), 1–16 (2023)
11. Tamir, T.S., Xiong, G., Li, Z., et al.: Traffic congestion prediction using decision tree, logistic regression and neural networks. *IFAC-Papers on Line* **53**(5), 512–517 (2021)



Development of an AI-Based International Trade Commodity Demand Forecasting Model

Wenle Qin¹, Yi Ling Wu², and Nan Chen³(✉)

¹ The Chinese University of HongKong, Shenzhen 518100, China

² Hong Kong Shue Yan University, Hong Kong 999077, China

³ University of Nottingham Ningbo, ZheJiang, NingBo 315000, China

nan.chen@niscsi.edu.cn

Abstract. Accurately predicting commodity demand in international trade is crucial for making import decisions. This paper proposes a deep learning-based demand forecasting model that integrates Long Short-Term Memory (LSTM) networks, attention mechanisms, and a hierarchical prediction framework. It effectively captures temporal dependencies, identifies key features, and predicts total and segmented demands in stages. Tested on a large-scale dataset covering 205 countries and regions and 1,268 commodities, the model demonstrates outstanding forecasting capabilities with a comprehensive Mean Absolute Percentage Error (MAPE) of 9.7%. Particularly, it achieves high accuracy of 6.8% and 7.2% for staple agricultural products like wheat and soybeans, significantly outperforming traditional machine learning models. This model provides robust technical support for enterprises and governments to formulate precise international trade decisions.

Keywords: International Trade · Demand Forecasting · Feature Engineering

1 Introduction

In the era of globalization, accurately predicting demand variations across countries and regions for various commodities is critical for shaping international trade policies and enterprise import decisions. Traditional statistical models and machine learning methods often struggle to comprehensively capture the complex factors influencing demand, failing to meet the precision requirements of real-world demand forecasting. Addressing this challenge, this paper proposes an innovative model that combines deep learning and attention mechanisms. It autonomously learns long-term dependency patterns in sequential data, identifies crucial features influencing predictions, and employs a hierarchical framework to predict total and segmented commodity demands in stages. Ultimately, it achieves high-precision trade demand forecasting, offering powerful technical support for formulating precise trade policies and decisions.

2 Theoretical Foundation

2.1 International Trade Theory

International trade theories provide a rational analytical framework for predicting commodity demand. Comparative advantage theory explains the fundamental reasons for division of labor and specialized production among different countries. According to this theory, each country should focus on producing products in which it has a comparative advantage to achieve optimal resource allocation. Therefore, when predicting a country's or region's demand for specific commodities, factors such as resource endowment, production costs, and technological level must be comprehensively considered [1]. Taking China's 2021 imports of agricultural products from the United States as an example, based on comparative advantage theory, China's significant imports of American soybeans, cotton, and other agricultural products are rational because these products exhibit clear comparative advantages in the United States. Relevant data indicates that in 2021, China imported approximately \$37.8 billion worth of agricultural products from the United States, with imports of soybeans and cotton reaching 32.8 million tons and 970,000 tons respectively, as shown in Table 1. This aligns with theoretical expectations.

Table 1. Major Agricultural Product Imports from the United States to China in 2021

Commodity	Import Volume
Soybeans	32.8 million tons
Cotton	970,000 tons

2.2 Commodity Demand Forecasting Models

Accurate prediction of commodity demand is crucial for formulating international trade strategies. Traditional demand forecasting models mainly include time series models, causal models, and composite models. Time series models analyze historical data to identify patterns such as cyclicity, trends, and seasonality in demand. Causal models start from various influencing factors to build models like multiple regressions. Composite models integrate multiple approaches to leverage their respective strengths [2]. For instance, to forecast China's demand for US soybeans, a multiple regression model can be constructed, incorporating variables such as population, domestic soybean production, feed demand, and substitute prices. The model expression is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

where Y represents China's demand for US soybeans, $X_1, X_2, \dots, X_n$ represent various influencing factors, and $\beta_0, \beta_1, \dots, \beta_n$ denote regression coefficients. By fitting historical data, the values of these coefficients can be estimated to forecast future demand. Compared to single time series models, this multiple regression-based forecasting model better captures the factors influencing demand, thereby improving prediction accuracy.

2.3 Application of Artificial Intelligence Technology

In the era of big data, traditional demand forecasting models exhibit certain shortcomings, such as inadequate capture of nonlinear relationships and difficulty in handling high-dimensional features. In contrast, artificial intelligence technology, especially deep learning models, possesses stronger capabilities in nonlinear fitting and feature learning, offering new possibilities for enhancing forecasting accuracy. For example, to predict global demand for US wheat, researchers have developed a deep learning model based on Long Short-Term Memory (LSTM) networks. This model not only learns traditional features such as wheat prices and substitute prices but also automatically extracts complex features such as climate and policies, comprehensively capturing diverse factors influencing demand. Experimental results show that the LSTM model achieves a Mean Absolute Percentage Error (MAPE) of 8.2%, significantly outperforming traditional autoregressive moving average models (13.9%) and multiple regression models (11.6%). As shown in Fig. 1, the LSTM model's predicted values (red line) closely align with actual values (blue line), indicating superior fitting performance [3]. This study demonstrates the broad application prospects of artificial intelligence technology in the field of international trade demand forecasting.

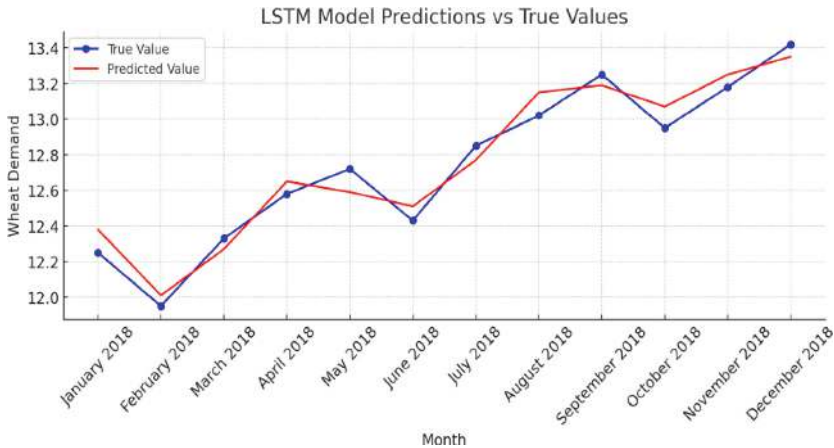


Fig. 1. Time Series of LSTM Model Forecasting US Wheat Demand

3 Model Construction

3.1 Data Collection and Preprocessing

To build a high-quality commodity demand forecasting model, comprehensive historical data is crucial as a foundation. We obtained annual import data from China to major countries and regions worldwide from the United Nations Comtrade database for the years 2000 to 2020, covering 310 commodity categories. These data reflect China's actual demand for various commodities over the past 20 years and serve as the basis for

model training. However, the raw data contains numerous missing values and outliers that require preprocessing. Initially, we employed interpolation methods to fill missing data, such as using the average of surrounding years' data. Subsequently, applying the 3-sigma rule, data points exceeding three times the standard deviation from the mean were identified as outliers and replaced with the median of corresponding data from other countries during the same period [4]. This step significantly enhanced the data quality and integrity. Additionally, due to substantial differences in data magnitudes across different commodity categories, we standardized the original data to mitigate potential biases from commodities with larger scales. Specifically, for each commodity, we calculated the mean μ and standard deviation σ over the 20-year period, transforming the original values x into standardized scores $z = (x - \mu) / \sigma$. Standardization ensures fairer model fitting across all commodities. Table 2 presents the standardized mean and standard deviation for several major commodities.

Table 2. Standardized Mean and Standard Deviation of Major Commodities

Commodity	Mean	Standard Deviation
Crude Oil	0	−1.03
Rice	−0.19	1.25
Cotton	0.09	−0.78

3.2 Feature Engineering

The performance of machine learning models largely depends on the quality of input features. Therefore, based on the original data, we enriched the feature set through feature engineering methods. Firstly, we encoded each country with a one-hot code based on economic development level, geographic location, and other features of import countries and regions, incorporating these categorical features into the model. Secondly, recognizing the significant cyclic and trend patterns in commodity demand, we constructed time-related features such as year and month, transforming them into periodic features using trigonometric encoding [5]. Additionally, for commodities like agricultural products sensitive to weather conditions, we collected climate data including annual rainfall and temperature for respective countries and regions as additional environmental features inputted into the model. Thirdly, we focused on price data for substitute and complementary goods of commodities, as changes in these prices often correlate closely with the demand for target commodities. For instance, soybean prices are strongly correlated with prices of soybean meal and show some substitution relationship with prices of other feed grains like corn. These price data of related commodities were integrated into the feature space. Finally, through techniques such as feature selection and dimensionality reduction, we identified over 200 features that contribute significantly to the model, filtering out redundant and noisy features to enhance the model's generalization performance. As depicted in Fig. 2, after implementing feature engineering, the root mean

square error on the test set significantly decreased, validating the effectiveness of feature engineering.

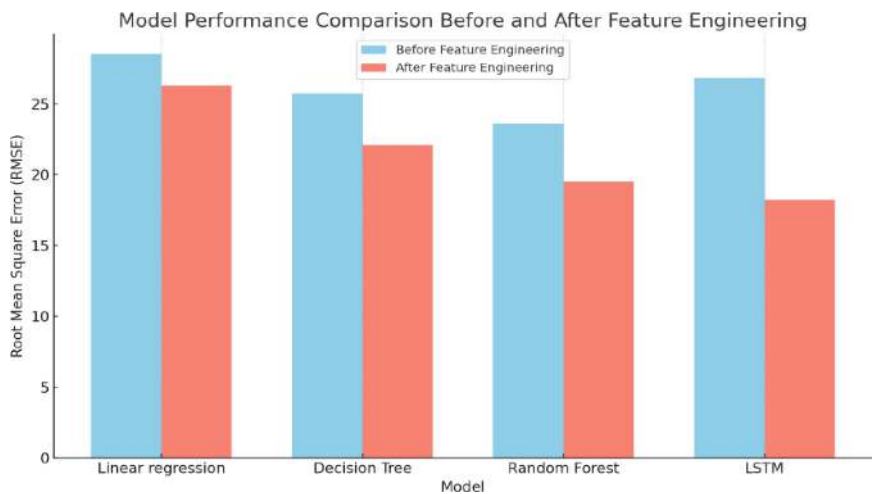


Fig. 2. Model Performance Comparison Before and After Feature Engineering

3.3 Model Selection and Training

Based on the aforementioned data and features, we experimented with various machine learning models, including linear regression, decision trees, random forests, and deep neural networks. Through cross-validation, we found that the deep neural network model, specifically the Long Short-Term Memory (LSTM) network, exhibited significant advantages and achieved the best fitting performance. Therefore, we ultimately selected an LSTM-based sequential model for forecasting commodity demand. LSTM is a specialized type of recurrent neural network (RNN) known for capturing long-term dependencies and periodic patterns in data, making it particularly suitable for time series prediction tasks. The model consists of three gate units: the forget gate controls the degree of forgetting old information, the input gate regulates the degree of acquiring new information, and the output gate determines the extent of outputting unit states. Overall, LSTM effectively addresses the vanishing or exploding gradient issues commonly encountered in traditional RNNs, thereby improving its ability to learn long-term dependencies [6]. We partitioned the past 20 years of data into three parts: a training set (2000–2016), a validation set (2017–2018), and a test set (2019–2020). Using the Adam optimization algorithm with root mean square error as the loss function, we conducted model training. After extensive experimentation and hyperparameter tuning, we determined the LSTM model’s configuration with 3 hidden layers and 128 units per layer. Additionally, to mitigate overfitting, we incorporated strategies such as Dropout layers and L2 regularization. Table 3 outlines the main hyperparameter settings of the model on the validation set.

Table 3. LSTM Model Hyperparameter Settings

Parameter	Value
Number of Hidden Layers	3 layers
Number of Units per Hidden Layer	128 units
Dropout Rate	0.2
L2 Regularization Coefficient	0.01

3.4 Model Evaluation and Optimization

To comprehensively assess the performance of the model, we utilized multiple metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). In the field of international trade, MAPE is widely used to evaluate the accuracy of demand forecasts, thus we adopted it as the primary evaluation metric. On the test set for 2019–2020, the baseline LSTM model achieved a MAPE of 8.1%, significantly outperforming traditional linear regression (16.3%) and decision tree (12.5%) models. As shown in Table 4, the LSTM model also demonstrated superior performance in MAE, RMSE, and other metrics [7]. This highlights the LSTM model’s strengths in capturing cyclical demands and handling long-term dependencies.

Table 4. Performance Metrics of Different Models on the Test Set

Model	MAE	RMSE	MAPE
Linear Regression	12.36	17.25	16.30%
Decision Tree	9.74	13.08	12.50%
LSTM	6.15	8.03	8.10%

Figure 3 further demonstrates the comparison between LSTM model predictions and actual values of China’s demand for US soybeans. It can be seen that the model effectively captures the trend of soybean demand over the years, thereby validating its effectiveness.

However, we also found that the baseline model exhibits some bias when predicting certain commodities, possibly due to unforeseen factors such as special events. To further optimize performance, we designed an LSTM-Attention model based on attention mechanisms, enabling the model to automatically learn to allocate weights to different features and enhance focus on key factors. Specifically, we introduced an attention layer after the LSTM’s output layer, which weights and sums the hidden states at each time step to generate the final output feature. The weight coefficients for each time step are determined by attention scores, calculated based on the similarity between the hidden states and context vectors. This allows the model to prioritize features that have a significant impact on predictions [8]. Additionally, we explored hierarchical prediction, where we first predict the total demand for primary categories of commodities and then predict

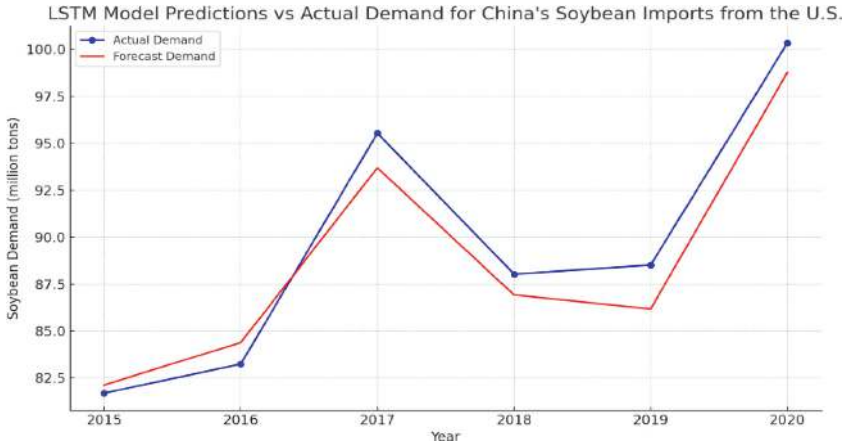


Fig. 3. LSTM Model Prediction of China's Demand for US Soybeans

specific demands for secondary categories. This hierarchical approach helps the model capture prediction patterns at different granularities, thereby improving accuracy. Ultimately, the LSTM-Attention hierarchical prediction model achieved a MAPE of 6.7% on the test set, demonstrating significant performance enhancement over the baseline model.

4 Experiment Results and Analysis

4.1 Experimental Setup

To comprehensively evaluate the AI-based international trade commodity demand forecasting model proposed in this study, we conducted extensive experiments using real-world data. The experimental data were sourced from the World Bank World Development Indicators database, covering annual import data from 2000 to 2019 for 205 countries and regions across 1268 commodity categories including agricultural products, energy products, and manufactured goods. The data were split into training, validation, and test sets in an 8:1:1 ratio. In the feature engineering phase, we constructed up to 312 features spanning time features, related commodity price features, climate and environmental features, national economic features, and more. Through feature selection, we retained 216 representative features for model input [9]. To assess the model's generalization ability, we trained and tested the models on different commodity categories. We established various baseline models for comparison, including Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Support Vector Regression (SVR), and conventional Recurrent Neural Networks (RNN). Additionally, we evaluated our proposed LSTM-Attention hierarchical prediction model as the primary model. During training, we employed the Adam optimization algorithm with a learning rate of 0.001, a batch size of 64, and 100 training epochs. For objective performance evaluation, we selected four commonly used regression task metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage

Error (MAPE), and R-squared (R^2) value. MAPE, reflecting the average relative error percentage between predicted and actual values, is considered a crucial metric in the field of trade demand forecasting. The comprehensive performance of different models on the test set is summarized in Table 5.

Table 5. Comprehensive Performance of Different Models on the Test Set

Model	MAE	RMSE	MAPE	R^2
LR	15.32	21.47	18.60%	0.672
DTR	12.85	17.93	15.80%	0.729
RFR	11.62	15.28	13.90%	0.783
SVR	10.74	14.05	13.10%	0.812
RNN	9.37	12.69	11.50%	0.845
LSTM-Attention Hierarchical	8.21	10.84	9.70%	0.883

4.2 Model Performance Evaluation

Table 5 presents the evaluation results of different models on the overall test set. It can be observed that the proposed LSTM-Attention hierarchical prediction model achieves the best performance across all metrics, achieving an excellent MAPE of 9.7%, significantly outperforming traditional linear models (18.6%), decision tree models (15.8%), and random forest models (13.9%). This fully validates the effectiveness of using deep learning models. It is noteworthy that even compared to other artificial intelligence models such as Support Vector Regression (13.1%) and conventional RNN (11.5%), the LSTM-Attention hierarchical prediction model also exhibits significant advantages. This is due to the LSTM’s outstanding ability to capture long-term dependency patterns and the contribution of attention mechanisms in automatically assigning feature weights and improving generalization performance through hierarchical prediction [10]. Figure 4 further illustrates the model’s performance across different commodity categories. It can be seen that the model achieves MAPE values within 10% for agricultural products, energy products, and metal products, demonstrating overall good predictive performance. Particularly in major agricultural products such as wheat (6.8%), soybeans (7.2%), and cotton (8.1%), the model achieves high levels of accuracy, providing strong support for import decision-making.

4.3 Results Analysis

The experimental results thoroughly demonstrate that the proposed LSTM-Attention hierarchical prediction model effectively captures complex factors influencing international trade commodity demand, achieving precise predictions. This is primarily attributed to the following aspects:(1)LSTM Network: Leveraging gate mechanisms

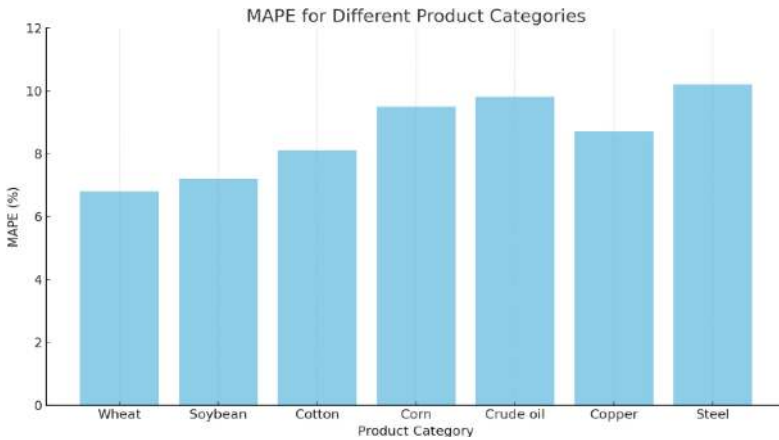


Fig. 4. MAPE Across Different Commodity Categories

and memory units, LSTM effectively learns long-term dependencies and periodic patterns in sequential data, thus accurately predicting changes in commodity demand trends.(2)Attention Mechanism: Endowing the model with the ability to automatically allocate feature weights, the attention mechanism allows it to focus more on critical features that significantly influence predictions, such as substitute prices and climate conditions.(3)Hierarchical Prediction Framework: By sequentially predicting total demand and segmented demands, the model enhances prediction accuracy for specific commodities and improves generalization capabilities.(4)Multidimensional Feature Engineering: Providing comprehensive input information including time features, related commodity prices, climate conditions, and national economic factors, the model enhances its representational capacity.(5)Large-Scale Real Data: Supported by extensive real-world data, the model learns objective patterns of trade demand, thereby enhancing its applicability in practical scenarios.

5 Conclusion

This paper proposes an artificial intelligence-based model for predicting international trade commodity demand, aimed at assisting enterprises and governments in making more precise import decisions. Integrating Long Short-Term Memory (LSTM) networks with attention mechanisms and a hierarchical prediction framework, the model effectively captures complex temporal dependencies and critical features influencing demand for agricultural products, energy products, industrial goods, and various other commodity categories. Tested on large-scale real data, the model achieves a low average absolute percentage error (MAPE) of 9.7%, with exceptional prediction accuracies of 6.8% for wheat and 7.2% for soybeans, significantly outperforming traditional and other artificial intelligence models. This robustly validates the model's outstanding performance in the field of international trade demand prediction.

References

1. Zhao, Q., Chen, J., Feng, X., et al.: A Novel Bézier LSTM Model: A Case Study in Corn Analysis. *Mathematics* **12**(15), 2227–7390 (2024). <https://doi.org/10.3390/math12152308>
2. Ameer, H.B., Boubaker, S., Ftiti, Z., et al.: Forecasting commodity prices: empirical evidence using deep learning tools. *Annals of Operations Research* **339**(1/2) (2024). <https://doi.org/10.1007/s10479-022-05076-6>
3. Wang, S., Wang, Y., Wang, J.: Spare parts demand forecasting method of modern enterprises based on digital twin model. *Int. J. Model. Simulat. Sci. Comp.* **13**(06) (2022). <https://doi.org/10.1142/S1793962322500453>
4. Dong, Y., Xiao, L., Wang, J., et al.: A time series attention mechanism based model for tourism demand forecasting. *Inf. Sci.* **628**, 269–290 (2023). <https://doi.org/10.1016/j.ins.2023.01.095>
5. Afia, A.E., Jebbor, S., Raddouane, C.: A preliminary study for selecting the appropriate AI-based forecasting model for hospital assets demand under disasters. *J. Humanit. Logist. Supp. Chain Manage.* **12**(1), 1–29 (2022). <https://doi.org/10.1108/JHLSCM-12-2020-0123>
6. Yong, L., Xiaojun, Y., Yi, L., et al.: A new emotion analysis fusion and complementary model based on online food reviews. *Comp. Electr. Eng.* **98**, 107679 (2022). <https://doi.org/10.1016/j.compeleceng.2022.107679>
7. Saravistha, D.B., Wyasa Putra, I.B., Sarjana, I.M., et al.: Digital trading forex: global trends and local impact. *J. Pharmaceut. Negat. Result* **14**(3) (2023). <https://doi.org/10.47750/pnr.2023.14.03.378>
8. Yuan, L., Hu, Y., Li, G., et al.: Biomass-based production of trimellitic and trimesic acids. *Green Energy & Environment* **9**(8), 1267–1278 (2023). <https://doi.org/10.1016/j.gee.2023.02.004>
9. Su, Q., Wang, H., Shi, X., et al.: Research on energy supply and demand forecast and carbon neutralization path based on grey-monarch butterfly optimization model. *J. Phys. Conference Series* **2247**(1), 012008 (2022). <https://doi.org/10.1088/1742-6596/2247/1/012008>
10. Chang, T.S.: Evaluation of an artificial intelligence project in the software industry based on fuzzy analytic hierarchy process and complex adaptive systems. *J. Enterp. Inf. Manag.* **36**(4), 879–905 (2023). <https://doi.org/10.1108/JEIM-02-2022-0056>



Research on Optimization Method of Product Purchasing Process Based on ASME Analysis——Take QH Company as an Example

Chunjie Li^(✉) and Jinfeng Wen

College of Business Administration, Guangdong Baiyun University, Guangzhou, China
40273706@qq.com

Abstract. With the rapid development of economy, purchasing management is playing a more and more important role in enterprises, and the continuous optimization of purchasing process is the key for enterprises to respond quickly to market demand. Taking QH Company as an example, this paper uses ASME analysis to analyze the procurement process, finds that there are problems of non-value-added and delay in the procurement process, and optimizes these problems. Through the comparison of the results before and after optimization, non-value-added activities and delayed activities have been significantly improved, thus shortening the procurement delivery time and improving work efficiency. The study of this method can be used as a reference for the procurement process optimization of other companies.

Keywords: Procurement Process · Process Optimization · ASME Analysis · E-commerce Procurement Platform · Supplier Management

1 Introduction

Optimizing product purchasing processes is vital for enhancing efficiency and competitiveness. QH Company seeks to streamline its purchasing operations. This research applies ASME Analysis to identify inefficiencies and propose improvements. we aim to develop strategies that will enhance QH Company's purchasing process, ultimately leading to cost savings and operational efficiency. This case study will demonstrate the effectiveness of ASME Analysis in optimizing product purchasing processes.

2 ASME Analysis Method

ASME analysis is a process analysis method formulated by the American Society of Mechanical Engineers. In this method, the activities in the process are divided into six types: value-added activities, non-value-added activities, inspection, delay, transportation and storage, so as to clearly and accurately judge the value of the business process. ASME analysis is a structured method that aims to clearly identify and describe each link

in the process. Value-added activities: activities that directly add value to a product or service. Non-value-added activities: activities that do not add value to a product or service, but are necessary to ensure the proper functioning of the process. Control/inspection activities: inspection or monitoring activities carried out to ensure quality or accuracy. Delivery activity: an activity that involves the transmission of information or products but does not add real value to the product or service. Delayed activities: activities that cause delays in the process, such as waiting for approval or processing. Among them, the judgment criteria of value-added activities and non-value-added activities are based on whether they can add value to the whole process or customers [1].

3 Current Situation of Product Purchasing Process of QH Company

QH company has established a complete procurement process management system, covering the whole process of new product introduction, supplier source search, purchase contract signing, purchase order execution and payment. In the whole procurement process, there is close cooperation and interaction within the enterprise (sales department, purchasing and logistics department, customization development department, finance section) and outside the enterprise (leasing warehouse, supplier). The overall procurement process of QH company is divided into planning project, developer design product instruction, purchase source search process, purchase order operation process, purchase order payment process, supplier performance scoring process and so on [2]. Figure 1 shows the procurement flow chart before QH company optimization.

4 ASME Analysis of the Whole Purchasing Process of QH Company Before Optimization

As shown in Table 1, there are 17 value-added activities, 32 non-value-added activities, 19 inspection activities, 14 transportation activities, 17 delayed activities and 3 storage activities in the whole procurement process.

At present, the purchase order payment process of QH Company is reconciliation at the beginning of the month, K3 system push purchase invoice, return, supplier seal and return, fill in payment application form, leader approval, submit payment application form and related information to financial accounting, financial accounting to check and check purchase invoice, financial cashier pays. With the exception of waiting time, the entire process takes time as shown in Table 2.

Table 2 shows that the whole payment process takes 121 min, which is mainly reflected in the reconciliation of the payment process and the leader's examination and approval.

5 Optimization Design of Purchasing Process of QH Company

This paper analyzes the procurement process optimization of QH Company from four aspects: procurement source search process, order operation process, purchase order payment process and supplier scoring performance process. ASME analysis is used to identify non-value-added activities and delay activities, and optimize these activities.

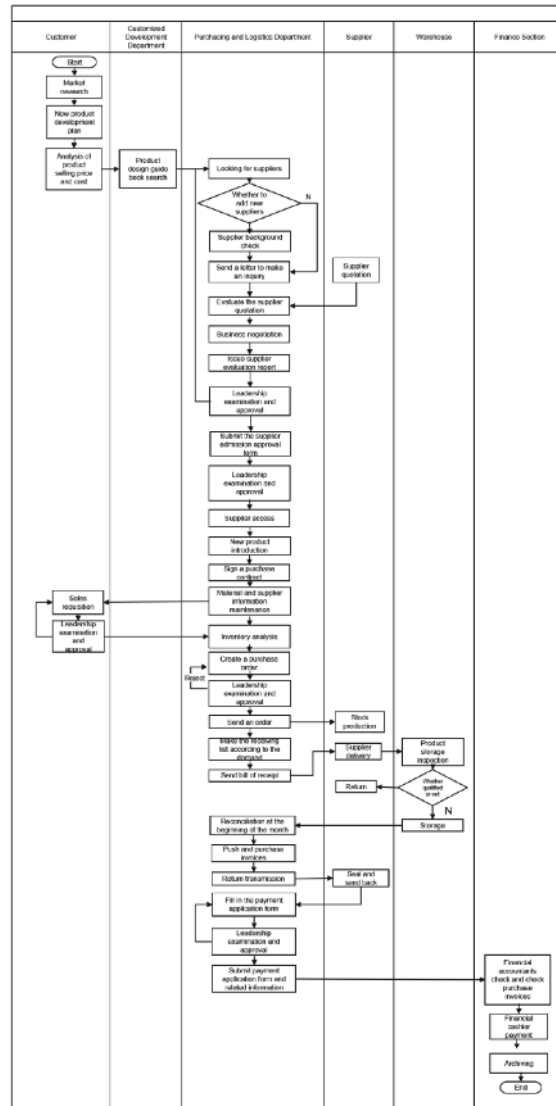


Fig. 1. Purchasing flow chart before optimization

5.1 Pre-procurement and Source-Seeking Work

In view of the fact that the previous submission of supplier access examination and approval form, supplier access examination and approval, and supplier access are completed offline, making use of the advantages of e-commerce procurement platform, let suppliers complete the information in e-commerce procurement platform, after the completion of submission and approval of e-commerce procurement platform, issue the signing of the procurement contract. Finally, the supplier information is synchronized

Table 1. ASME analysis of the whole purchasing process before optimization

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
1	Set up a project	Market research	✓						Planning specialist
2		New product development plan	✓						Planning specialist
3		Analysis of product selling price and cost	✓		✓				Planning specialist
4		Product design instruction		✓	✓				Development specialist
5	Procurement source search process	Looking for suppliers	✓		✓		✓		Procurement specialist
6		Supplier background check	✓		✓				Procurement specialist
7		Send a letter to make an inquiry		✓		✓			Procurement specialist
8		Supplier quotation		✓		✓	✓		Supplier
9		Evaluate the supplier quotation	✓		✓				Procurement specialist
10		Business negotiation	✓						Procurement specialist
11		Issue supplier evaluation report	✓		✓				Procurement, development
12		Leadership examination and approval		✓	✓		✓		Relevant leaders
13		Submit the supplier admission approval form		✓		✓			Procurement specialist
14		Leadership examination and approval		✓	✓		✓		Relevant leaders
15		Supplier access		✓					Procurement specialist
16		New product introduction		✓					Procurement specialist
17		Sign a purchase contract	✓			✓	✓		Procurement specialist
18		Material and supplier information maintenance		✓	✓			✓	Procurement specialist
19	Order operation process	Sales requisition	✓						Sales specialist

(continued)

Table 1. (continued)

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
20		Leadership examination and approval		✓	✓		✓		Relevant leaders
21		Inventory analysis	✓		✓		✓		Procurement specialist
22		Create a purchase order		✓					Procurement specialist
23		Leadership examination and approval		✓	✓		✓		Relevant leaders
24		Send an order		✓		✓			Procurement specialist
25		Supplier stock production	✓				✓		Supplier
26		Make the receiving list according to the demand		✓			✓		Procurement specialist
27		Send bill of receipt		✓		✓			Procurement specialist
28		Supplier delivery	✓			✓			Supplier
29		Product storage inspection		✓	✓		✓		Warehouse specialist
30		Storage/return according to inspection results	✓			✓			Warehouse specialist
31	Purchase order payment process	Reconciliation at the beginning of the month		✓	✓		✓		Procurement specialist
32		K3 push purchase invoice		✓					Procurement specialist
33		Return transmission		✓		✓			Procurement specialist
34		The supplier stamped and sent back		✓		✓			Supplier
35		Fill in the payment application form		✓					Procurement specialist
36		Leadership examination and approval		✓	✓		✓		Relevant leaders
37		Submit payment application form and related information		✓		✓			Procurement specialist

(continued)

Table 1. (continued)

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
38		Financial accountants check and check purchase invoices		✓	✓				Financial accounting
39		Financial cashier payment		✓					Financial cashier
40		Paper filing		✓				✓	Financial cashier
41	Supplier performance scoring process	Data recording		✓			✓	✓	Procurement specialist
42		Procurement initiates supplier rating assessment		✓		✓			Procurement specialist
43		Relevant personnel fill in the supplier score form		✓	✓	✓	✓		Purchasing, quality
44		Leadership examination and approval		✓	✓		✓		Relevant leaders
45		Send a score sheet		✓		✓			Procurement specialist
46		Supplier feedback and communication	✓						Supplier, Procurement specialist
47		Work together to develop improvement plans	✓						Supplier, Procurement specialist
48		Examination and approval of improvement plan		✓	✓		✓		Relevant leaders
49		Follow up the improvement plan	✓						Procurement specialist
		Total	17	32	19	14	17	3	

from the e-commerce procurement platform to Kingdee K3 system to complete the maintenance of material and supplier information. Among them, after the new product introduction work is confirmed by the supplier, we can start the preparation work. Complete the supplier's mass production notice and send it to the supplier before the completion of the contract. The optimized procurement source search process is shown in Fig. 2.

Among them, the release of demand, supplier registration and product quotation, release results, supplier access, the signing of procurement contracts are completed on the e-commerce procurement platform, and the maintenance of material and supplier

Table 2. The time of each step before the optimization of the purchase order payment process is as follows

Serial number	Activity	Time(Minute)	Operator
1	Reconciliation at the beginning of the month	60	Procurement specialist
2	K3 push purchase invoice	2	Procurement specialist
3	K3 push purchase invoice	2	Procurement specialist
4	The supplier stamped and sent back	10	Supplier
5	Fill in the payment application form	5	Procurement specialist
6	Leadership examination and approval	20	Relevant leaders
7	Submit payment application form and related information	2	Procurement specialist
8	Financial accountants check and check purchase invoices	10	Financial accounting
9	Financial cashier payment	5	Financial cashier
10	Archiving	5	Financial cashier
	Total	121	

information is synchronized to Kingdee K3 system from the e-commerce procurement platform. The rest of the process is completed online or offline.

5.2 Citing E-Commerce Purchasing Platform

When the product is stored in Kingdee K3 system, the incoming information will be synchronized to the purchase order of the e-commerce purchase platform. The optimized order operation process is shown in Fig. 3.

Among them, the creation of sales orders, purchase requirements, purchase orders, inventory early warning trigger supplier delivery, supplier delivery, product storage/return are completed on the e-commerce procurement platform. When a formal purchase order is generated, the purchase order is automatically sent to the supplier and synchronized to the Kingdee K3 system to generate the purchase order. When the Kingdee K3 system enters and returns, the information will be synchronized to the purchase order corresponding to the e-commerce purchase platform [3].

5.3 Automatic Payment for Non-Prepaid Purchase Orders

Through the introduction of e-commerce procurement platform and Kingdee K3 system integration to optimize the purchase order payment process, for all purchase order payments should follow the whole process approval optimization. Due to the automatic

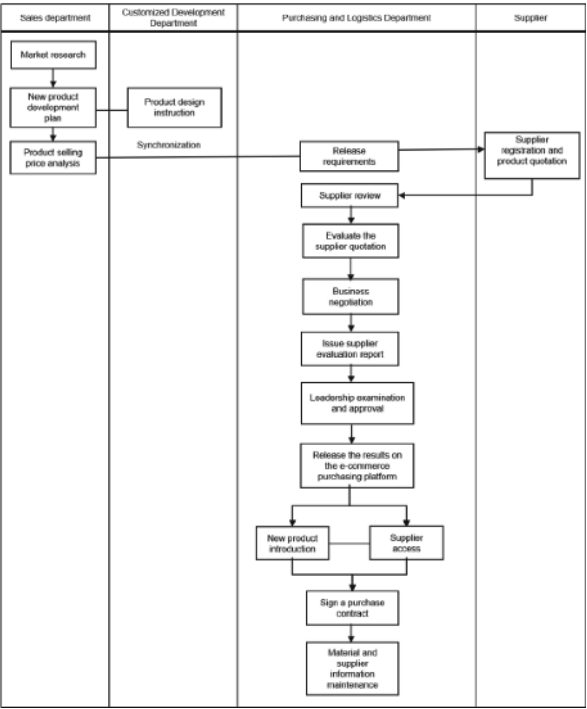


Fig. 2. Flow chart after optimization of purchasing source search

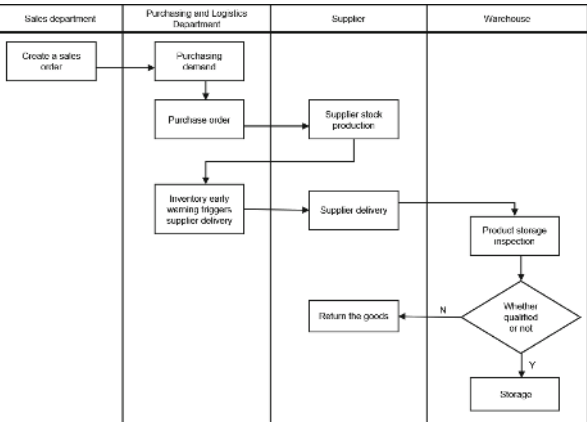


Fig. 3. Flowchart after order operation optimization

reconciliation between the Kingdee system and the e-commerce procurement platform at the beginning of the month, the probability of payment errors is relatively small. You can cancel the two processes of filling in the payment application form and leading the examination and approval process for the purchase order payment process of

non-fixed assets in the enterprise, so as to save time and improve efficiency. Submit the purchase invoice provided by the supplier to the Finance Department, which checks and checks the invoice and the cashier's payment, in which the financial payment can be made according to the invoice time on the purchase invoice and the account period of the supplier on the Kingdee system is equal to the payment time. For the purchase order payment process with advance payment, in order to avoid errors in the amount, the original payment process is retained. After the payment is completed, the e-commerce procurement platform automatically transmits the payment return order to the supplier so that the supplier can verify the payment. Finally, automatic electronic filing is realized through the e-commerce procurement platform [4]. The procurement payment process optimized through the e-commerce procurement platform is expected to take time as shown in Table 3.

Table 3. The time of each link after the optimization of non-prepaid purchase order is as follows

Serial number	Activity	Time(Minute)	Operator
1	Automatic reconciliation	10	Procurement specialist
2	K3 push purchase invoice	2	Procurement specialist
3	Return transmission	2	Procurement specialist
4	The supplier stamped and sent back	10	Supplier
5	Submit purchase invoice	2	Procurement specialist
6	Financial accountants check and check purchase invoices	10	Financial accounting
7	Financial cashier payment	5	Financial cashier
8	Electronic filing	2	Financial cashier
	Total	43	

The non-prepayment and payment process shown in Table 3 has been reduced from 121 min to 43 min, resulting in a significant reduction in payment time and improved efficiency. Procurement specialists can be allowed to do more value-added activities [5]. The optimized non-prepaid purchase order payment process is shown in Fig. 4.

5.4 Improve the Supplier Management Process

In view of the imperfect supplier management process, the supplier performance scoring process is analyzed, and it is found that the supplier performance scoring work is time-consuming and formal, and there is no dynamic assessment of the supplier [6]. Through the use of e-commerce procurement platform to optimize the supplier performance scoring process to achieve supplier scoring dynamic assessment. For the products delivered to the warehouse, by setting the supplier performance evaluation index and weight in the e-commerce procurement platform, after the completion of each order purchase, the

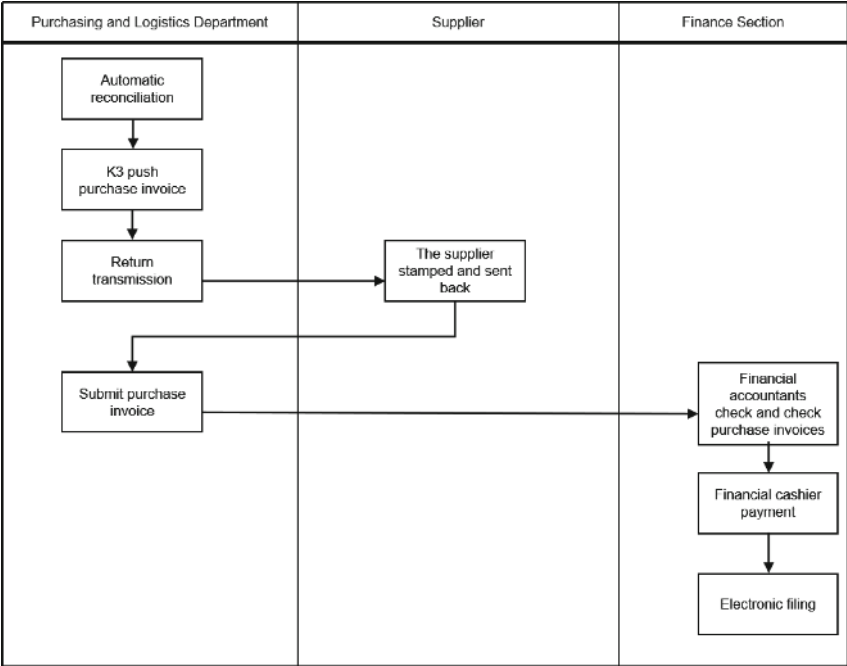


Fig. 4. Flowchart after payment optimization of non-prepaid purchase order

system automatically pushes the supplier performance score form to the warehouse, procurement, quality and other related personnel, the relevant personnel fill in the supplier score form in time. After filling in, it will be automatically pushed to the relevant ministers for approval, and the supplier scoring results will be automatically returned to the corresponding suppliers after the approval of the e-commerce procurement platform. Suppliers and company leaders can log on to the e-commerce procurement platform at any time to view supplier performance results [7]. For the suppliers with poor performance, the purchasing and logistics department will take the lead to work out the improvement plan with the development and quality of the customized development department, which will be approved by the leader and sent to the supplier on the e-commerce purchasing platform. And continue to follow up the supplier improvement results, if the supplier refuses to cooperate and the enthusiasm is poor, reduce the order in the future cooperation, terminate part of the cooperation, the new product introduction does not consider the supplier, let the supplier pay attention to the improvement plan put forward by our company. If the supplier performs well, incentives can be made. Priority will be given to cooperation in increasing orders and introducing new products in the future. The supplier performance scoring process for delivery to the warehouse is optimized as shown in Fig. 5.

For the product that the supplier sends directly to the customer, because the product does not pass through the warehouse, the purchasing specialist cannot obtain the product quality, at this time, the supplier performance scoring process can be completed by the customer. Set a reasonable delivery time limit in the e-commerce procurement platform,

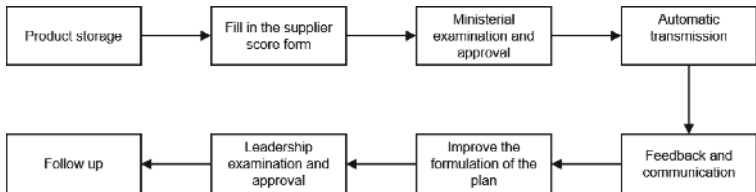


Fig. 5. Flowchart after optimization of supplier performance score for delivery to warehouse

after the customer signs for the product, the e-commerce procurement platform can automatically push the evaluation information of the product. Star rating and text evaluation are carried out in quality, timeliness, logistics and other aspects [8]. The Purchasing specialist regularly exports customer evaluation data and summarizes them. Send the results of supplier performance evaluation on the e-commerce purchasing platform. The next process is consistent with the process that the supplier sends to the warehouse. The supplier performance scoring process of the supplier sending the product directly to the customer is optimized as shown in Fig. 6.

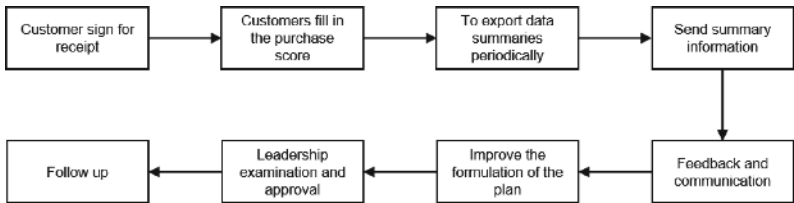


Fig. 6. Flow chart after optimization of supplier performance score for straight hair products

Finally, the after-sales service evaluation process is added in the e-commerce procurement platform, and the after-sales link of the products should also be communicated to suppliers and customers. Therefore, after-sales service is also an index to assess suppliers, through the e-commerce procurement platform after-sales service section, set the corresponding indicators and weights.

6 ASME Analysis of the Whole Purchasing Process of QH Company After Optimization

It can be seen from Table 4 that there are 14 value-added activities, 19 non-value-added activities, 10 inspection activities, 15 delivery activities, 4 delayed activities and 2 storage activities in the whole process (except supplier scoring process) after procurement optimization [7]. Before optimization, there are 14 value-added activities, 26 non-value-added activities, 16 inspection activities, 11 transportation activities, 13 delayed activities and 2 storage activities in the whole procurement process (except supplier scoring process). A column chart is used to show the changes of the six activities in the whole procurement process before and after optimization, as shown in Fig. 7.

Table 4. ASME analysis of optimized procurement process

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
1	Set up a project	Market research	✓						Planning specialist
2		New product development plan/product design instruction	✓	✓	✓				Planning, Development
3	Procurement source search process	Product price cost analysis/looking for suppliers	✓		✓				Planning, Procurement
4		Supplier registration and product quotation	✓			✓			Supplier
5		Supplier review	✓		✓				review panel
6		Evaluate the supplier quotation	✓		✓				Procurement specialist
7		Business negotiation	✓						Procurement specialist
8		Issue supplier evaluation report	✓		✓				Procurement, Development
9		Leadership examination and approval		✓	✓		✓		Relevant leaders
10		Publish the results		✓					Procurement specialist
11		Supplier admission/new product introduction/Purchasing specialist		✓		✓			Procurement specialist
12		Sign a purchase contract	✓			✓	✓		Procurement specialist
13		Material and supplier information maintenance		✓				✓	Procurement specialist
14	Order operation process	Create a sales order	✓			✓			Sales specialist
15		Ministerial examination and approval		✓	✓	✓			Relevant leaders
16		Purchasing demand		✓		✓			Procurement specialist
17		Purchase order		✓		✓			Procurement specialist
18		Ministerial examination and approval		✓	✓		✓		Relevant leaders
19		Automatic transmission		✓		✓			Procurement specialist
20		Supplier stock production	✓						Supplier
21		Inventory warning prompt for suppliers to deliver goods	✓			✓			Procurement specialist

(continued)

Table 4. (continued)

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
22		Supplier delivery	✓			✓			Supplier
23		Product storage inspection		✓	✓				Warehouse specialist
24		Storage/return according to inspection results	✓			✓			Warehouse specialist
25	Purchase order payment process	Automatic reconciliation		✓		✓			Procurement specialist
26		K3 push purchase invoice		✓					Procurement specialist
27		Return transmission		✓		✓			Procurement specialist
28		The supplier stamped and sent back		✓		✓			Supplier
29		Submit purchase invoice		✓		✓			Procurement specialist
30		Financial accountants check and check purchase invoices		✓	✓				Financial accounting
31		Financial cashier payment		✓					Financial cashier
32		Electronic filing		✓				✓	Financial cashier
33	Direct delivery product performance score	Customer sign for receipt	✓						Customer
34		Customers fill in the purchase score	✓						Customer
35		Export data summaries periodically		✓					Procurement specialist
36		Send summary information		✓		✓			Procurement specialist
37		Feedback and communication	✓						Supplier, Procurement
38		Improve the formulation of the plan	✓						Supplier, Procurement
39		Leadership examination and approval		✓		✓	✓		Relevant leaders
40		Follow up	✓						Procurement specialist
41	Delivery to warehouse performance score after sale	Product storage	✓			✓			Warehouse specialist
42		Fill in the supplier score form		✓					Procurement, Quality
43		Ministerial examination and approval		✓		✓	✓		Relevant leaders
44		Automatic transmission		✓		✓			Procurement specialist

(continued)

Table 4. (continued)

Serial number	Process flow	Activity	Value-added activities	Non value-added activities	Check	Transport	Delay	Storage	Operator
45		Feedback and communication	✓						Supplier, Procurement
46		Improve the formulation of the plan	✓						Supplier, Procurement
47		Leadership examination and approval		✓		✓	✓		Relevant leaders
48		Follow up	✓						Procurement specialist
49		Customer feedback problem	✓						Customer
50		The supplier receives and resolves	✓			✓			Supplier
51		Customers fill in after-sales evaluation	✓						Customer

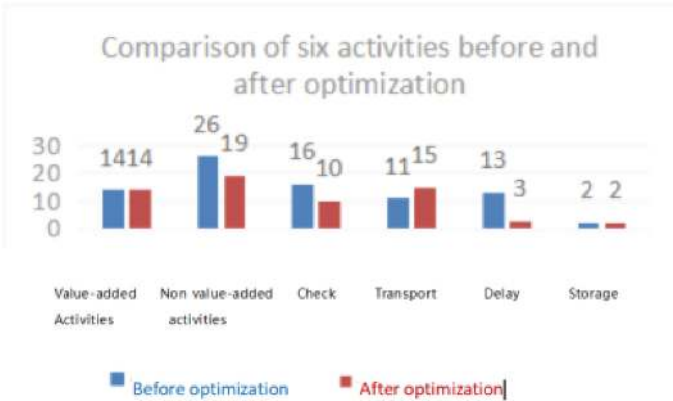


Fig. 7. Comparison of six activities before and after optimization

As can be seen in Fig. 7, value-added activities remain unchanged, but some steps can be combined to improve work efficiency at the same time, non-value-added, inspection and delay activities are reduced, manual operations are reduced, and work efficiency is improved. Although the transportation activities have increased, most of them are automatic transportation, which reduces the manual operation, and the storage activities remain unchanged before and after optimization [9].

7 Summary

In this paper, through the analysis of the current situation of the procurement process of QH Company, starting with the procurement source search process, order operation process, purchase order payment process, supplier scoring performance process, using

ASME analysis, analyze which are non-value-added activities, which are delayed activities, and optimize these activities. Procurement process optimization is a process of continuous optimization, which can help enterprises adapt to the changes of the market environment and improve market competitiveness and adaptability.

References

1. Reflection on the Process Reengineering of University Procurement Business from the Perspective of Internal Control - Based on ASME Method for Process Reengineering of J University He Li; Shen Yihong; Zheng Qifang, Chief Accountant of China, (04) (2021)
2. Weihuan: Research on Purchasing process Optimization of D Company. Harbin Institute of Technology (2022)
3. Research on supplier Management Optimization of Tang Yinhua A Company. East China normal University (2023)
4. Zhuo, J.B.: Research on purchasing process optimization strategy of H company. Xi'an University of Architectural Science and Technology (2022)
5. Research on procurement process optimization of MRO general products of Wanjiang D Company. Chang'an University (2023)
6. Lin, X.: Tr Technology Company Purchasing Process Optimization Research. Shandong University (2023)
7. Irwin, J.: AI could innovate purchasing processes. *Automotive News* **98**(7109) (2023)
8. Michela, G., Federico, C., Antonella, M., et al.: The role of artificial intelligence in the procurement process: State of the art and research agenda. *J. Purchas. Supp. Manage.* **29**(2) (2023)
9. Althabatah, A., Yaqot, M., Menezes, B., et al.: Transformative procurement trends: integrating industry 4.0 technologies for enhanced procurement processes. *Logistics* **7**(3) (2023)



Digital Transformation and Sustainable Development of Guangdong's Manufacturing Sector

Guiling Fan^(✉)

Guangdong University of Science and Technology, Dongguan, China
fanguiling@gdust.edu.cn

Abstract. Sustainable development is crucial for realizing China's modernization goals. In the digital economy era, enterprise digitalization serves as a key driver for sustainable growth. This study examines data from Guangdong's A-share listed manufacturing companies spanning 2012 to 2022, investigating the effects of digital transformation on sustainable development, a vital component of the national economic framework. The findings demonstrate that digital transformation positively influences the sustainable growth of Guangdong's manufacturing sector. Additionally, the impact varies based on regional differences and firm-specific attributes. Targeted recommendations are provided in light of these findings. This research offers strategic sights for advancing digital transformation in Guangdong, a core region of China's industrial landscape.

Keywords: Digital Transformation · Guangdong · Manufacturing Companies · Sustainable Development · Total Factor Productivity

1 Introduction

As the digital economy continues to advance rapidly, industries are increasingly adopting digital transformation strategies to meet evolving market needs and enhance their competitiveness. Recognizing the strategic importance of this shift, enterprise digitalization has emerged as a key priority at the national level. In January 2022, the State Council unveiled the “14th Five-Year Plan for Digital Economy Development,” underscoring the urgency of “accelerating enterprise digitalization and upgrading” as well as “deepening digital transformation across key sectors.” This policy framework has elevated the discourse on enterprise digitalization within academic and industrial circles.

The 19th National Congress of the Communist Party of China (CPC) emphasized that the nation's economic development has entered a new era, marked by the pursuit of sustainability and high-quality growth. Later, the 20th National Congress reaffirmed sustainable development as a top priority in the comprehensive endeavor to build a modern socialist country. Achieving sustainable development is thus an essential pathway toward realizing Chinese-style modernization. The manufacturing sector, a vital pillar of the national economy, underpins China's strength, forms the foundation of its industrial

framework, and acts as a key engine for sustainable growth. As such, advancing the manufacturing sector's development with a focus on quality has become a critical area of academic inquiry.

Digital transformation is widely regarded as a crucial approach for businesses aiming to achieve long-term, high-quality growth. Nevertheless, its impacts differ substantially across sectors (Wang Fujun et al., 2023) [1]. The 19th National Congress emphasized the necessity of promoting advancements in high-value manufacturing and integrating cutting-edge technologies into the broader economy. Guangdong, as a prominent manufacturing hub in China, plays a significant role in this transformation process. This research explores how digital transformation influences the sustainable progress of manufacturing enterprises in Guangdong, analyzing data from the region's A-share listed firms between 2012 and 2022. The results are intended to provide actionable strategies and recommendations to further enhance the region's digitalization efforts in manufacturing.

2 Literature Review and Research Hypothesis

The resource-based theory suggests that variations in enterprise development stem from the heterogeneity of their internal resources, enabling certain firms to gain competitive advantages within the same industry [2]. Among these resources, data assets are increasingly recognized as a critical factor in enterprise growth. Digital transformation has significantly reshaped the foundational concepts and operational models of the manufacturing sector, offering fresh perspectives for its advancement and empowering the sector's sustainable development [3]. Empirical studies have shown that implementing digital transformation strategies positively influences the manufacturing industry's progress (Liu Mingyang et al., 2024; Qi Yudong et al., 2020) [4, 5], with noticeable incremental benefits (Liu Hedong et al., 2023; Liu Xinxin et al., 2021) [6, 7].

The essence of sustainable enterprise development lies in enhancing total factor productivity (Gao Yue, 2023) [8]. A growing body of research demonstrates that digital transformation markedly boosts total factor productivity (Zhao Chenyu et al., 2021; Tu Xinyue et al., 2022) [9, 10]. Notably, its impact strengthens over time, making it a key driver for productivity improvement in the digital economy era (Tu Xinyue et al., 2022; Zhao Chenyu et al., 2021).

Digital transformation facilitates optimized resource allocation, enhances supply efficiency [11], lowers production costs, and boosts productivity. It also elevates product quality [9], fosters technological innovation, improves management efficiency, advances intelligent production processes, and maximizes economies of scale [10]. These improvements collectively drive sustainable enterprise development.

Hypothesis: Digital transformation positively influences the sustainable development of manufacturing companies in Guangdong.

3 Study Design

3.1 Study Sample

This research focuses on manufacturing firms listed in Guangdong's A-share market between 2012 and 2022 as the dataset. The sample underwent the following screening steps: first, enterprises with significant missing data for key variables were excluded; second, firms classified as PT, ST, *ST, or those undergoing delisting or consolidation processes were removed; third, continuous variables were trimmed within the 99% confidence interval to eliminate outliers. Following these steps, an unbalanced panel dataset was constructed, comprising 4905 observations from 734 listed Guangdong manufacturing firms. The most of the data source for this analysis was the Cathay Pacific database.

3.2 Variable Selection

Dependent Variables. This study utilizes total factor productivity (TFP) as a representative metric to evaluate the manufacturing sector's progress toward high-quality and sustainable development. TFP is derived through calculations employing methods such as LP, OP, OLS, and GMM, in accordance with established formulas.

$$TFC = Y/AL^{\alpha} K^{\beta} \quad (1)$$

Independent Variable. Digital Transformation serves as the independent variable in this study. Building upon the approach proposed by Wu et al. (2021) [12], Python crawler technology and text analysis techniques are employed to construct a digital transformation index for enterprises.

Control Variables. This study incorporates 13 control variables informed by prior research. Moreover, time-fixed effects are included in the analysis. Table 1 presents a detailed overview of all variables along with the methods used for their calculation.

3.3 Model Setting

To begin, certain variables were transformed using logarithmic processing. Based on the Hausman statistic test results, it was determined that the association between company digital transformation and the long-term growth of manufacturing firms in Guangdong is best analyzed using a fixed-effects model.

$$(\hat{\beta}_{FE} - \hat{\beta}_{RE})' [Var(\hat{\beta}_{FE} - \hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \rightarrow^d \chi^2(K) \quad (2)$$

$$TFP\_LP_{it} = \beta_0 + \beta_1 DT_{it} + \beta_j Controls_{it} + \sum Year + \varepsilon_{it} \quad (3)$$

Table 1. Variable Definition.

Types	Names	Variable Symbols	Variable Definition
Dependent Variable	Manufacturing Companies' Sustainable Development	TFP_LP	TFP calculated through the LP method
Independent Variable	Digital Transformation	DT	Calculated based on feature thesaurus using text method mining method
Control Variable	Company Age	AGE	Year of listing plus 1, taking the natural logarithm
	Enterprise Size	SIZE	Total assets of enterprises take natural logarithm
	Enterprise Attributes	SOE	State-owned enterprises take 1, otherwise 0
	Level of Competition	MARKET	Gross profit margin of enterprise sales
	Government Subsidies	SUBSIDY	Government subsidies in non-recurring profit and loss/revenue from main business of the firm
	Innovation Ability	INNOVATION	The total number of patents authorized by the enterprise is adjusted by adding 1 before applying the natural logarithm
	Profitability	PROFIT	Enterprise return on net assets
	Cash Flow	CASH	Operating cash flow to total assets ratio
	Growth	GROWTH	Growth rate of operating revenue
	Operating Capability	MANAGE	Total Assets Turnover Ratio
	Financial Leverage	LEVERAGE	Ratio of total liabilities to total assets
	Shareholding Concentration	SHARE	Ownership percentage of the top ten shareholders

(continued)

Table 1. (continued)

Types	Names	Variable Symbols	Variable Definition
	Board Size	BOARD	Number of directors taking the natural logarithm
	Year	YEAR	Annual Dummy Variables

4 Study Results and Analysis

4.1 Scatterplot Analysis

A scatter plot was generated using STATA software to examine the association between DT and the long-term growth of Guangdong manufacturing firms (Fig. 1). The plot suggests a positive linear correlation between the two variables, providing preliminary evidence that digital transformation positively influences the sustainable development of companies.

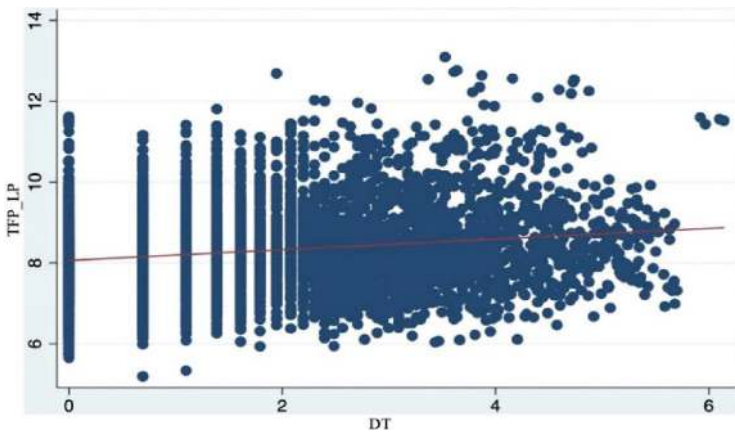


Fig. 1. Scatterplot of DT and TFP_LP.

4.2 Descriptive Statistical Analysis

Descriptive statistics were computed using the following formula, with the results presented in Table 2.

$$Mean_{variable} = variable_1 + variable_n + \cdots + variable_n/n \tag{4}$$

$$\overline{var} = \sum_{i=1}^n var_i/n \tag{5}$$

$$S_{var} = \sqrt{\frac{\sum_{i=1}^n (var_i - \overline{var})^2}{n - 1}} \quad (6)$$

As shown in Table 2, there are notable variations in the development levels across the sample firms in Guangdong's manufacturing sector, along with considerable discrepancies in their digital transformation efforts. These differences in both technological advancement and development quality provide valuable perspectives on how digital change influences business growth. Such variation is crucial for conducting subsequent regression and econometric analyses, which will provide deeper insights into these dynamics.

Table 2. Descriptive analysis of primary variables.

Variables	Obs	Means	Std. Dev	Min	Max
TFP_LP	4,905	8.316325	1.052488	5.193659	13.09613
DT	4,905	1.918221	1.437345	0	6.139884
AGE	4,905	2.039263	0.7772359	0.6931472	3.465736
SIZE	4,905	22.12915	1.290734	17.6413	28.29301
SOE	4,905	0.2340469	0.4234448	0	1
MARKET	4,905	0.2949256	0.1680426	-0.6494305	0.9499939
SUBSIDY	4,905	0.0067878	0.0123738	0	0.0751945
INNOVATION	4,905	0.5185403	1.344737	0	7.021084
PROFIT	4,905	0.0610382	0.5217373	-4.23363	33.49516
CASH	4,905	0.048785	0.0750012	-0.5561684	0.6837366
GROWTH	4,905	0.267228	3.092155	-0.9069014	167.6457
MANAGE	4,905	0.6682284	0.4953923	0.0126236	7.133285
LEVERAGE	4,905	3.597244	4.61577	0.5111016	125.4842
SHARE	4,905	58.78095	14.98467	3.5876	101.16
BOARD	4,905	2.083281	0.2018872	1.609438	2.890372

4.3 Benchmark Regression Analysis

To examine how digital transformation influences the long-term growth of manufacturing firms in Guangdong, a fixed-effects regression was conducted with the aid of Stata software. The findings are summarized in Table 3. In the first column, the model shows the results based solely on the primary explanatory variable, while the second column includes additional control variables for a more comprehensive analysis.

The results indicate that, digital transformation consistently demonstrates a strong and positive impact on the long-term growth of Guangdong's manufacturing industry,

with significance at the 1% level. This finding supports the hypothesis put forth in this study.

Table 3. Benchmark regression results.

Variables	TFP_LP (1)	TFP_LP (2)
DT	0.1087664*** (11.99)	0.0256223*** (4.88)
AGE		−0.0927057*** (−4.45)
SIZE		0.5784472*** (59.81)
SOE		−0.007205 (−0.27)
MARKET		0.0259907 (0.50)
SUBSIDY		−2.19316*** (−4.96)
INNOVATION		−0.00519 (−1.46)
PROFIT		0.004381 (0.64)
CASH		0.3294399*** (5.63)
GROWTH		−0.0113167*** (−9.62)
MANAGE		1.090961*** (66.20)
LEVERAGE		0.0005976 (0.55)
SHARE		−0.0007253 (−1.23)
BOARD		−0.0837845** (−2.30)
YEAR	control	control
_cons	7.714255*** (299.42)	−4.982071*** (−23.86)
N	4905	4905
Adj-R ²	0.2803	0.7734

4.4 Robustness Testing

Endogeneity Issues. This research addresses potential endogeneity issues, including reverse causality and omitted variable bias, which could influence the relationship between digital transformation and enterprise growth. To consider the delayed effects of digital transformation on the long-term outcomes of enterprises [10, 13], we re-evaluate the primary variable, digital transformation, using a fixed-effects panel model with one-period and two-period lag structures. As shown in Table 4, the coefficients for digital transformation remain highly significant at the 1% level, even after accounting for these time lags. These results further substantiate the conclusion that digital transformation plays a vital role in driving the progress of Guangdong's manufacturing sector.

Table 4. Endogeneity test results.

Variable	TFP_LP (1)	TFP_LP (2)
DT _{t-1}	0.0356994*** (6.54)	
DT _{t-2}		0.0280684*** (4.82)
Controls	control	control
YEAR	control	control
_cons	−4.707314*** (−20.33)	−4.187527*** (−15.73)
Adj-R ²	0.7695	0.7500

Replacement of Key Variables. To validate the robustness of the results, key variables in the analysis are replaced. First, the dependent variable is altered. This study recalculates the total factor productivity (TFP) of enterprises using the OP, OLS, and GMM methods. The outcomes, displayed in columns (1) to (3) of Table 5, consistently demonstrate that regardless of the estimation technique, TFP, which integrates digital transformation indicators, continues to exert a significant positive influence on the sustainable development of manufacturing firms in Guangdong. Subsequently, the independent variables are adjusted to gain deeper insights into the mechanisms by which digital transformation affects enterprise performance. At this stage, “big data utilization” and “frequency of digital technology application” are employed as proxies for digitalization levels. The results, presented in columns (4) and (5) of Table 5, reveal that these factors significantly enhance the sustainable development of Guangdong's manufacturing enterprises. These findings remain consistent across different models, confirming the robust impact of digital transformation.

In summary, the results from Table 5 demonstrate that regardless of changes to either the dependent or independent variables, the conclusions remain stable. This further supports the central hypothesis of the study: Digital transformation has a meaningful and positive impact on advancing the sustainable development of Guangdong's manufacturing sector.

Table 5. Replacement of the calibers of calculation of key variables.

Variable	Replace Dependent Variables			Replace Independent Variables	
	TFP_OP	TFP_OLS	TFP_GMM	Big Data	Digital technology applications
	(1)	(2)	(3)	(4)	(5)
DT	0.022333*** (4.22)	0.0185492*** (4.39)	0.0265555*** (4.40)	0.002583 *** (4.56)	0.002462*** (5.73)
Controls	control	control	control	control	control
YEAR	control	control	control	control	control
N	4905	4905	4905	4905	4905
Adj-R ²	0.7155	0.8767	0.6294	0.7732	0.7732

5 Heterogeneity Analysis

To explore how the impact of digital transformation on the sustainable development of manufacturing firms in Guangdong differs by region and specific company factors, this study performs subgroup regressions based on geographical and enterprise characteristics. The results of this analysis are detailed in Table 6.

As shown in columns (1) to (4) of Table 6, there are noticeable regional differences; however, digital transformation consistently has a positive effect on the sustainable development of manufacturing companies throughout Guangdong. The subgroup regression results show that digitalization positively influences sustainable development in enterprises throughout central, northern, western, southern, and eastern Guangdong. However, the strength of this influence differs significantly by region. Notably, the less economically developed northern and western parts of Guangdong experience a stronger impact, this may be attributed to their relatively lagging digital infrastructure and the slower pace of digital adoption. In contrast, the more developed area, southern, and eastern regions have already made considerable progress in adopting digital technologies, so the additional benefits of digital transformation are less pronounced. In these regions, where the digitalization process is more advanced, the potential for further gains is smaller. On the other hand, the northern and western areas, where digitalization is still in its early stages, benefit more from the implementation of digital technologies, significantly accelerating their progress toward sustainable development.

Furthermore, the regional disparities in digital transformation’s impact are not merely due to the speed of adoption but also the local economic conditions and the varying levels of technological infrastructure. In regions with less mature digital ecosystems, manufacturing enterprises are more likely to experience a significant leap in productivity and quality improvement as digital technologies fill gaps in the local industrial value chains. In contrast, in more advanced regions, where digital infrastructure is already well-established, the room for substantial improvement through digital transformation is more limited, thus resulting in a relatively smaller additional impact.

In columns (5) to (6) of Table 6, the examination of ownership structure's influence on the connection between digital transformation and sustainable development reveals a generally positive effect for both state-owned and privately owned enterprises in Guangdong. However, the impact is more significant in privately owned enterprises. This disparity can be attributed to the more adaptable organizational structures and management approaches typically found in these companies, which are better positioned to integrate and utilize new digital technologies. Privately owned enterprises often demonstrate greater flexibility and are quicker to implement innovative strategies and technologies, which enhances their operational efficiency, productivity, and overall sustainable development. In contrast, state-owned enterprises may encounter more bureaucratic hurdles, slower decision-making processes, and more rigid organizational systems, which can delay their digital adoption. Therefore, while both sectors benefit from digital transformation, the effect is typically stronger and more immediate in the privately owned sector, where adaptability and innovation are more easily embraced.

Table 6. Results of heterogeneity analysis.

Variable	Regional Attributes				Enterprise Attributes	
	Central Guangdong	Northern and Western Guangdong	Southern Guangdong	Eastern Guangdong	State-Owned Enterprises	Non-State-Owned Enterprises
	1	2	3	4	5	6
DT	0.080066*** (10.63)	0.3971859*** (9.23)	0.0538761*** (10.00)	0.0643831*** (4.49)	0.0540177*** (5.02)	0.0723542*** (16.33)
Controls	control	control	control	control	control	control
_cons	-5.498258*** (-21.51)	-0.970727 (-0.66)	-5.334354*** (-33.46)	-6.465936*** (-10.85)	-5.602676*** (-19.38)	-5.761444*** (-40.18)
N	1138	75	3183	357	1148	3757
Adj-R ²	0.8794	0.8793	0.8501	0.8658	0.8451	0.8527

6 Conclusions and Recommendations

This research empirically demonstrates the beneficial impact of digital transformation on promoting sustainable development within manufacturing companies in Guangdong. The findings are robust and consistent. The analysis further shows that manufacturing enterprises in less economically developed regions, such as northern and western Guangdong, gain more significant advantages from digital transformation. Additionally, the study reveals that privately owned manufacturing enterprises experience a more substantial positive effect from digital transformation than their state-owned counterparts.

- (1) **Emphasizing the Importance of Digitalization.** Manufacturing enterprises in Guangdong must recognize the essential role digitalization plays in improving their competitiveness and responding to market dynamics. It is crucial to incorporate

digitalization into the enterprise's core strategic framework to drive sustainable development.

- (2) Formulating a Digital Transformation Strategy. Enterprises should develop a clear and actionable digital transformation strategy that guides their decision-making and operations throughout the transformation process.
- (3) Building Digital Culture and Organizational Capabilities. It is important for Guangdong manufacturing enterprises to nurture a digital culture that encourages people to embrace and adapt to digital transformation. Additionally, enterprises should enhance their organizational digital capabilities by training employees, recruiting digital talent, and fostering cross-departmental collaboration.
- (4) Investing in Digital Infrastructure in Less Developed Regions. Increased investment in digital technology and infrastructure is crucial, particularly for manufacturing enterprises in the northern and western regions of Guangdong, where the marginal returns of digital transformation can be maximized.
- (5) Promoting the Digital Advancement of Privately Owned Enterprises. Policy-makers should prioritize supporting the digital enhancement and modernization of privately owned manufacturing enterprises, as these companies show greater potential to capitalize on digitalization. This, in turn, will foster the overall sustainable development of the sector.

This study is limited by space and concentrates exclusively on how digital transformation influences the sustainable development of manufacturing firms in Guangdong. Future research could investigate the specific mechanisms by which digitalization affects business performance and analyze how these effects differ among various types of enterprises and regions.

Acknowledgment. This paper presents findings from the university-level 'Innovation and Strengthening University Project' (research category) project (GKY-2022CQPY-5) and the general humanities and social sciences project (GKY-2022KYYBW-8).

References

1. Fujun, W., Hao, W.: Digital transformation, internal control and high-quality development of enterprises. *Friends of Accounting* (24), 100–106 (2023)
2. Bertrand, M.: How much should we trust differences-in-differences estimates. *Risk Management and Insurance Review* **119**(1) (2004)
3. Yong, J.: Digital economy enabling manufacturing transformation: from value remodeling to value creation. *The Economist* (6) (2020)
4. Liu, M., Wan, Y., Tang, X., et al.: Study on the impact of digital transformation on high-quality development of manufacturing industry in the context of "one belt and one road"-based on the perspective of upgrading industrial structure and digital trade. *Xinjiang Social Science* 1–20 (2024). <http://kns.cnki.net/kcms/detail/65.1211.F.20240110.1512.002.html>
5. Qi, Y., Liu, H.: Research on production factor attributes of data and its marketized allocation mechanism under digital economy. *Economy Vertical and Horizontal* (11), 63–76 (2020)
6. Liu, H., Ji, R.: Research on the mechanism and effect of digital economy to promote industrial structure upgrading. *Sci. Technol. Progr. Countermeas.* (1), 61–70 (2023)

7. Liu, X., Hui, N.: Research on the impact of digital economy on the high-quality development of China's manufacturing industry. *Economic system reform* (5), 92–98 (2021)
8. Yue, G., Yuetang, W., Yunyang, J.: Digital transformation empowers enterprise high-quality development-an analysis based on the perspective of urban heterogeneity. *Urban Issues* **06**, 95–103 (2023)
9. Chenyu, Z., Wenchun, W., Xuesong, L.: How digital transformation affects enterprise total factor productivity. *Finance and Trade Economics* **42**(07), 114–129 (2021)
10. Xinyu, T., Xiaoling, Y.: Digital transformation, knowledge spillovers and firm total factor productivity - empirical evidence from listed manufacturing companies. *Ind. Econ. Res.* **02**, 43–56 (2022)
11. Ping, Q., Weihui, S.: Digital transformation, innovation development and high-quality development of manufacturing value chain. *J. f Guangdong Univ. Finan. Econ.* **39**(01), 4–18 (2024)
12. Wu, F., Hu, H., Lin, H., et al.: Corporate digital transformation and capital market performance - Empirical evidence from stock liquidity. *Manage. World* (7), 130–144+10 (2021)
13. Zhichao, L., Jianxin, W.: The impact of digital transformation on the high-quality development of enterprises. *Statist. Decis. Mak.* **39**(22), 162–167 (2023)



Research on the Application of Information Technology in Modern Enterprise Economic Management

Fanling Kong¹(✉) and Qiwen Zhang²

¹ Guangdong University of Science and Technology, Dongguan 523129, Guangdong, China
kongfunling@126.com

² Northeast Agricultural University, Harbin 150030, Heilongjiang, China

Abstract. Aiming at the problems existing in the technical and economic management of construction enterprises, a modern enterprise economic management system based on information technology is designed. This system can effectively meet the technical and economic management needs of modern enterprises, such as engineering quality management, overall planning, and staff training. The test results show that the system can support 200 people operating online simultaneously with rapid feedback, meeting the daily management needs. It plays a crucial role in improving the efficiency of enterprise economic management and promoting the digital development of construction enterprises.

Keywords: Economic Management · Information Technology · Modern Enterprise

1 Introduction

Many construction companies today utilize standalone cost management software, resulting in critical data such as project quotations and cost calculations being isolated within individual computers, making it difficult to compare, locate, and share information. Furthermore, the coordinated management of construction technical regulations, economic rules, software usage, and project quality has become complex due to insufficient sharing. Against the backdrop of rapid information technology development, the technical level of enterprise economic management has significantly improved, to a certain extent promoting the healthy development of construction enterprises. Currently, there are relatively many domestic and international studies on the application of information technology in enterprise economic management. Some studies have proposed a modernized intelligent economic management framework in analyzing enterprise economic management systems and built enterprise economic management systems based on big data technology [1]. Other studies have further enhanced the intelligence level of economic management systems by combining big data technology with cloud computing technology, laying the foundation for building dynamic economic management systems [2].

Based on this, this paper proposes an information technology-based economic management system for construction enterprises, aiming to provide a reference for achieving dynamic economic management in construction enterprises and building upon the existing research.

2 System Overall Design

2.1 System Architecture

The innovative B/S architecture, based on the traditional C/S two-tier structure, adopts a three-tier architecture model, which is suitable for building Web applications. In the B/S architecture, users can access the system through a Web browser on their devices to connect to the server. The server-side connects to the database system, utilizing web technologies to achieve system openness, simplicity, and ease of later functional expansion and maintenance [3]. Common Web browsers include Netscape Navigator, while mainstream database products include Oracle and SQL Server. The advantage of the B/S architecture lies in the fact that system updates and functional enhancements can be performed solely on the server side, without the need to modify the client side, thus simplifying the entire software development and operational maintenance process. The B/S architecture consists of three main components: the data access layer, the business logic layer, and the presentation layer. The presentation layer, which is the user-interactive frontend, employs technologies such as HTML, CSS, and JavaScript to create an intuitive user interface for information display and user interaction. The business logic processing layer handles the core functionality of the application, developed using the C# programming language. The data access layer manages data storage and retrieval. This layered design makes the system more modular and easier to manage and maintain. A traditional B/S architecture system diagram is shown in Fig. 1.

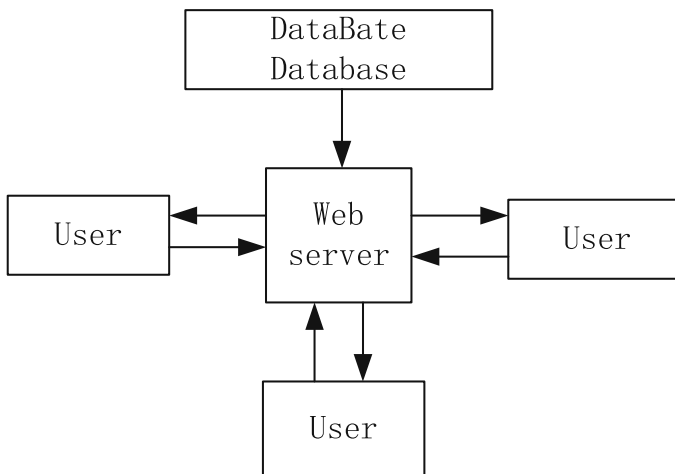


Fig. 1. System Architecture

2.2 Cloud Platform Virtual Server Deployment

Utilizing cloud storage technology, the system realizes remote data preservation and establishes a comprehensive data management and storage solution on the cloud, implementing the installation of a cloud-based office system. This strategy not only reduces maintenance costs but also optimizes system operation efficiency [4]. In the initial stage, the technical and economic department needs to obtain storage space on the cloud platform to facilitate system deployment, and then make adjustments according to business expansion. Additionally, the system adopts “virtual hosting” technology to generate customized virtual web servers on the cloud to meet the specific needs of different departments. This configuration significantly simplifies the data backup and server upgrade processes, enhancing the system’s adaptability and maintenance convenience. The system deployment is illustrated in Fig. 2.

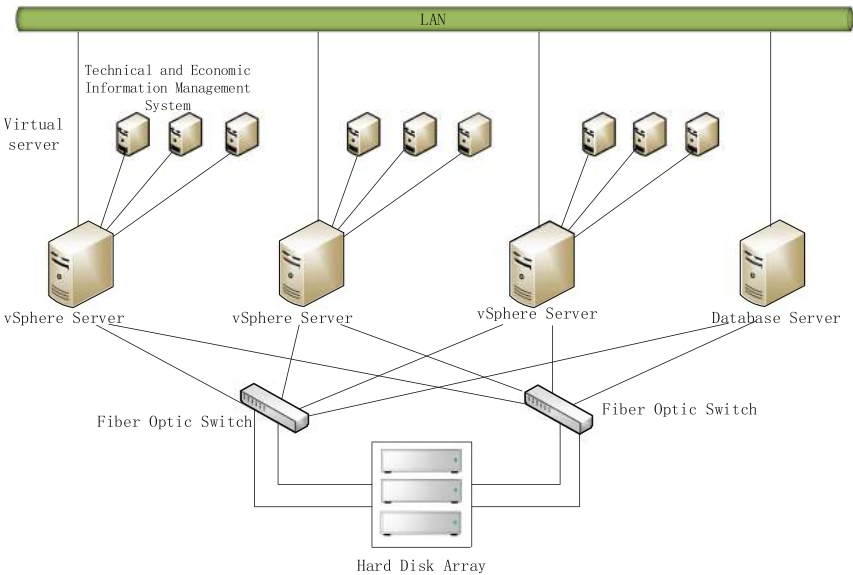


Fig. 2. System Deployment Architecture Diagram

3 Functional Design and Implementation

3.1 Overall Functional Design

Currently, cost management software for construction enterprises is mostly operated as stand-alone applications, limited to use on a single device, resulting in project budgets and related information being stored only on individual employees’ computers, adversely affecting information sharing and data query efficiency [5]. Therefore, an information management system should be established to facilitate the unified management of all

documents and project data, enabling functions such as data comparison, rapid search, self-management, and maintenance. By centralizing and dispersing cost information, the efficiency of retrieval and comparison is improved. At the same time, the system will also organize and classify the files held by various departments and individuals within the company uniformly, and set up corresponding permission management to ensure the security and orderly use of files, making file management more standardized. The system design will be divided into six main modules: login verification module, integrated business processing module, project management module, technical data management module, quality control module, and training and education module. These modules will be further subdivided to jointly ensure that the economic and technical information system can fully perform its intended functions [6]. Through these modules, the technical and economic management department can centrally collect cost data from all individuals and departments within the company and manage them on a unified platform, greatly simplifying the process of data query and analysis. The basic framework modules of the technical and economic management information system are shown in Fig. 3.

3.2 Functional Design and Implementation

(1) Login Module

The login module provides two login methods for users and administrators. After the user submits the account and password, the server will receive a verification request sent by the system. If the verification is successful, the user will be directed to the main interface; if not, an error message will be displayed [7]. This module employs MD5 encryption to protect data security. Once the server receives the request, the system backend will conduct in-depth verification of the user's identity based on preset rules and assign corresponding permissions accordingly. The implementation process of the login module is shown in Fig. 4.

(2) Design of the Integrated Management Module

After logging into the system, users will enter the homepage interface named "Integrated Control," which serves as a toolbox integrating multiple commonly used functional modules [8]. Users can compose new notifications, view unread notifications, search notification history, manage files, view data statistics, and access technical and economic documents. The process of composing a notification is similar to writing a letter. Users need to first access the "Integrated Control" module through the main menu and select the "Draft Notification" option. After clicking the "Create" button, the writing process begins. Notifications typically include a title, type (e.g., notification or announcement), recipient (specific department or group), receiving members (specific individuals), and the main content. For notifications that are no longer needed or contain errors, users can select the corresponding notification and click "Delete" to remove it. If modifications are required, users can double-click on the notification or select it and enter the "Details" page for editing, similar to editing an existing email. After composing or modifying a notification and preparing to publish it, users can find the corresponding entry in the notification list, click "Immediate Publish," and the notification will be

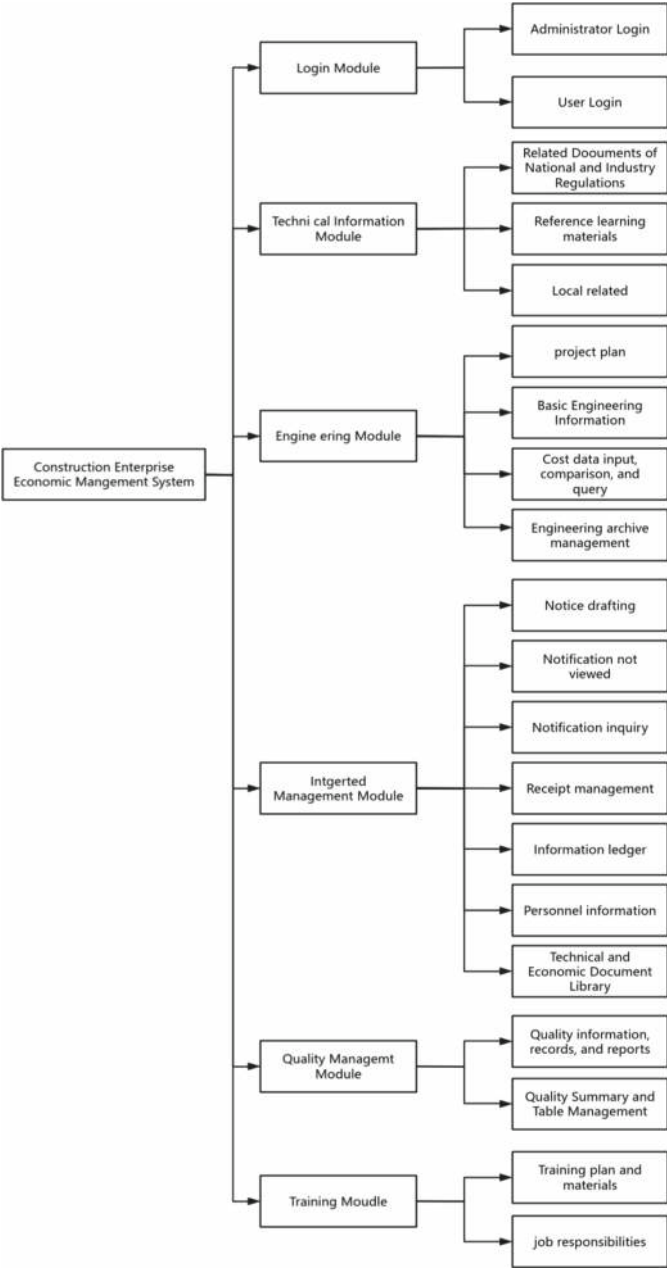


Fig. 3. System Functional Structure

immediately sent to the predetermined recipients. The process design of the Integrated Management Module is shown in Fig. 5.

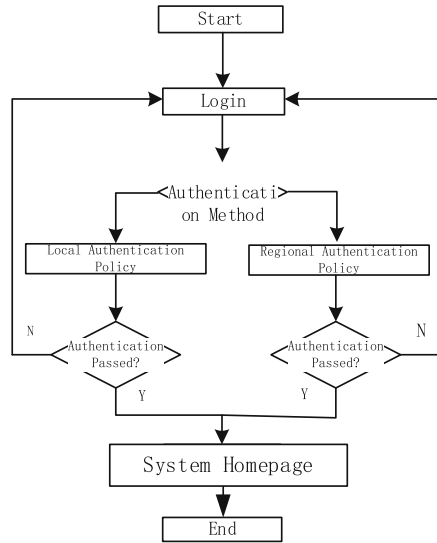


Fig. 4. Flowchart of the Implementation of the Login Module

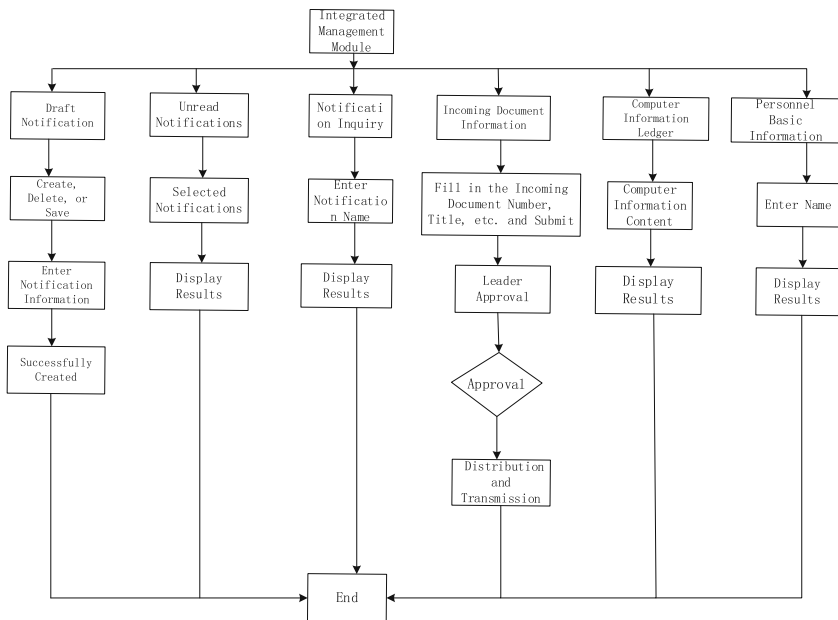


Fig. 5. Flowchart Corresponding to the Integrated Management Module

(3) Technical Documentation Module

The Technical Management Module functions like a library, containing local quotas, relevant regulations, documents, and reference materials, which users can read and save [9]. Editing permissions for critical technical documents are limited to administrators to ensure information updates. Regular users can only read and have no authority to modify or delete, effectively ensuring the integrity and accuracy of the materials. The code for cloud download of technical documents is as follows:

```
function km_download (repo,gid,uType){
var grid=Ext.getCmp(gid);
var recs=grid.getSelectionModel().getSelections();
if(recs.length==0)return false;
var folders=[],files=[];
for(vari=0;i<recs.length;++i){
if(recs[i].get('category')=='Folder'){
folders. push (recs[i]. get (' rid'));
}else {
files. push (recs[i].get (' rid'));
}
}
window.                                     open('/WebGroup/KM/km_download.
aspx?repo='+repo&folders='+folders.join(',')+'&files='+files.join(',')+'&uType='+
uType);
}
```

(4) Engineering Module

After logging into the system, users will enter the main page, which includes a customized functional area for engineering projects, resembling a toolbox. This area covers functions such as basic project information, work plans, cost data, comparative analysis, and file management. The basic project information constitutes an online information database that is updated in real-time, with all data originating from the company's internal Electronic Document Management System (EDMS), ensuring the timeliness and integrity of the information. System administrators strictly manage access and operation permissions for project information, and only personnel with comprehensive management responsibilities are authorized to add, modify, or delete information according to company regulations. Engineering information is closely linked to cost data [10]. The operational process is as follows: Users select "Engineering Projects" from the main menu, then click "Basic Information" upon entering. From the list of project records, users can choose a specific project and click "View Details" to access detailed information about that project, which is updated in real-time to reflect the latest status.

(5) Quality Management Module

After logging into the system, users will enter the main page, which includes a dedicated area for quality management that encompasses two core functions: quality information recording and analysis. The quality information recording function documents

routine quality data such as the name, source, content, and resolution of the information, with related files attached as attachments. These files need to be updated regularly to ensure the timeliness of the information. The operational process for the Quality Management Module is as follows: Users access “Quality Management” through the main menu and select “Quality Information” to enter [11]. On this interface, users can browse the list of quality records and click “View Details” to view the specific details of each record, akin to opening an envelope to read the enclosed document, with all displayed content being the latest data. The process design of the Quality Management Module is shown in Fig. 6.

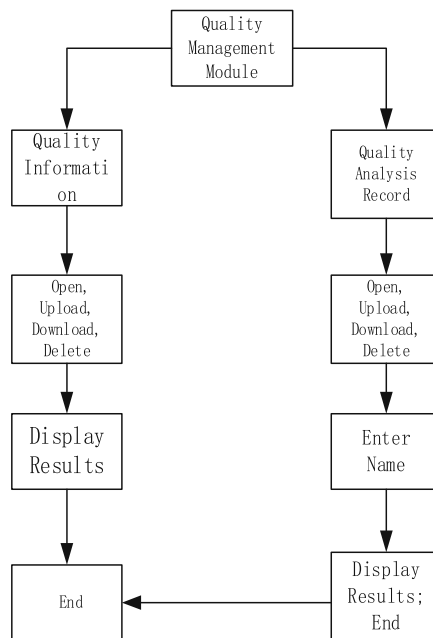


Fig. 6. Flowchart Corresponding to the Quality Management Module

(6) Design of the Education and Training Module

After logging into the system, users will enter the main page and see a dedicated section for education and training, which is divided into two parts: annual training schedule and training materials. The annual training schedule keeps track of the training plans for the entire year, with each plan having a unique number and name for easy identification. The corresponding original plan documents are attached to each record, similar to attaching photos, and need to be updated regularly to maintain the effectiveness of the information. Users can access “Education and Training” through the main menu, enter the “Training Plan” section, and browse the entire list of plans [12]. By clicking the “View” button next to any plan, users can access detailed information, similar to

flipping through a book, with all displayed information being the latest version. The process design of the Education and Training Module is shown in Fig. 7.

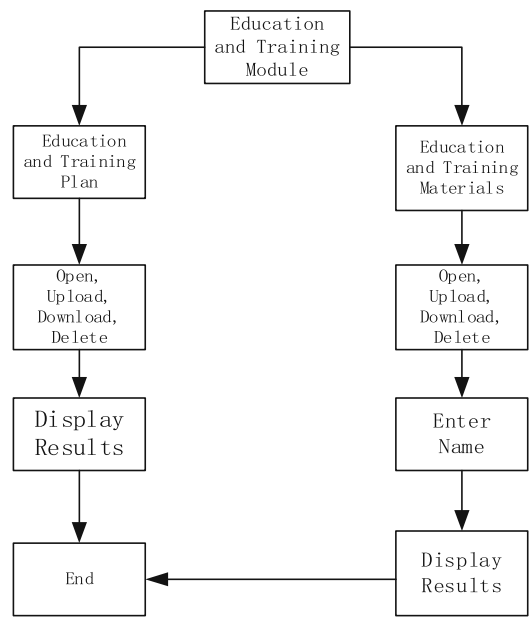


Fig. 7. Flowchart of the Education and Training Module

4 System Testing

4.1 Functional Testing

Functional testing is conducted using JMeter, a functional testing tool designed to verify if every part of a software system functions correctly. By simulating both normal and exceptional usage scenarios, the system's functional blocks are checked individually to validate if they meet user requirements and achieve the desired outcomes [13]. This process helps identify and fix issues, ensuring software quality and reliable delivery to users.

After multiple functional tests, the system interface did not experience any issues such as inaccessible pages, lags, or error messages. The functionality was consistent with the expected results. For instance, the test results for the login function and user permission verification are presented in Tables 1 and 2.

4.2 Performance Testing

Construction enterprises conduct comprehensive performance testing on their technical and economic management systems to ensure stable operation under various conditions.

Table 1. Test Results for Login Function

Function Description	Login	
Test Case Purpose	To verify if the login function is functioning properly through testing	
Preconditions	Using local login method, with username: User, and password: 666888	
Operation Type	Expectation	Actual Result
Correct username and password	Enter the system's homepage	Consistent
Incorrect password	Prompt for username or password error, and exit the login module	Consistent
Incorrect username	Prompt for username or password error, and exit the login module	Consistent
User logout	User logs out and returns to the login page	Consistent

Table 2. Test Results for User Permission Verification Function

Function Description:	User Permissions	
Test Case Purpose	To test if a user has the corresponding permissions	
Preconditions	Using regional login method, with username: User, and password: 666888	
Operation Type	Expectation	Actual Result
Clicking on an authorized operation node	Redirect to the relevant interface	Consistent
Clicking on an unauthorized operation node	Prompt for user permission error	Consistent

The testing is divided into two main parts: stress testing and performance testing. Stress testing is typically conducted under resource-constrained conditions, such as insufficient equipment resources, limited memory, tight hard disk space, and slow network speed. It primarily involves gradually increasing the system's workload, such as the amount of information, files, or requests, to test the maximum capacity of the system. The goal is to observe at which point the system experiences performance degradation or failures, thus determining its limit of operational intensity [14]. Performance testing focuses on the system's performance under different environmental conditions, including processing speed, response time, and other key indicators. Load testing, as a part of performance testing, evaluates the system's efficiency under high load conditions by assigning a large number of tasks to it, testing if it can maintain efficient operation despite the increased workload. This helps identify potential weak points in the system and understand its performance at critical points, such as the maximum number of concurrent users or the amount of information it can support [15]. Furthermore, to

accommodate the company’s current and future needs, the testing includes gradually increasing the number of concurrent users on the system, from 20 to 200 or more. This measures the system’s performance under different user loads, including response speed and CPU utilization, enabling enterprises to accurately assess whether the system can meet their current and future operational requirements. The test results are shown in Fig. 8.

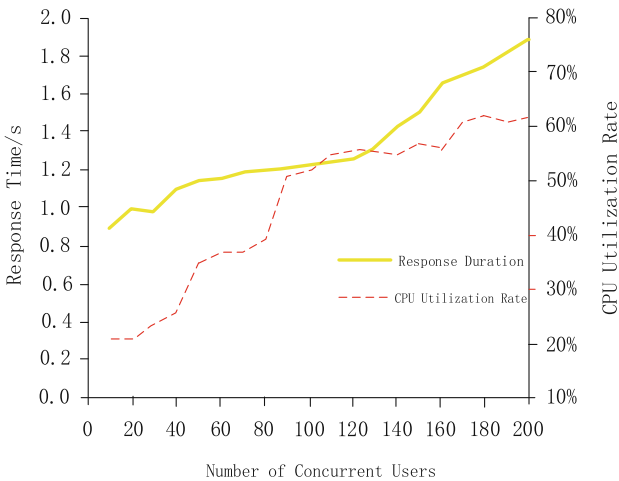


Fig. 8. Performance Test Results

As can be seen from Fig. 8, as the number of users continues to increase, the response time of the system also experiences a slight extension, and the processor occupancy rate gradually rises. For instance, when the number of users reaches 200, the system’s response time remains within two seconds, and the processor occupancy rate is only 61%. At this stage, the system still operates normally. Therefore, this system can fully satisfy the management needs of the technical and economic management departments of ordinary enterprises.

5 Conclusion

In summary, this article conducts a study on the technical and economic management system of construction enterprises. This system not only addresses the challenges faced by construction enterprises in terms of technology, economy, and data management, but also enhances the centralization of information management and operational convenience through online processing workflows. After thorough stress testing and load testing, the system has proven to meet preset performance standards, providing strong technical support for the daily operations of construction enterprises. It is expected that this system will continue to demonstrate its value in practical applications, contributing to the sustainable development and innovation of the construction industry.

Acknowledgement. This research was supported by Foundation: Ph. D. Research Start-up Fund of Guangdong University of Science & Technology (Grant No. GKY-2024BSQDW-1).

References

1. Wang, T.: Role of big data based on information system in management of agricultural economy. *Front. Soc. Sci. Technol.* **5**(13), 176–200 (2023)
2. Cui, Y.: Research on effective measures of economic management in construction enterprises in the new era. *China Collect. Econ.* **1**, 68–71 (2023)
3. Cai, J.: Thoughts on the informatization of economic management in construction enterprises. *Construct. Sci. Technol.* **7**(1), 76–77, 80 (2023)
4. Gong, F.: Improvement path of enterprise economic management in the era of big data. *Front. Soc. Sci. Technol.* **5**(9), 58–65 (2023)
5. Jianzeng, Y.: Reflections on the optimal integration of economic strategy and human resource management in construction enterprises. *Finance Econ.* **19**, 174–176 (2023)
6. Hong, C.: A brief discussion on effective measures for economic management of construction enterprises in the new era. *J. Hubei Open Vocat. College* **36**(21), 122–124 (2023)
7. Alexey, S., Vladimir, B., Eduard, B.: Mathematical support to make managerial decisions in the socio-economic systems of public transport enterprises. *MATEC Web Conf.* **4**(10), 341 (2021)
8. Fushuai, C.: Research on financial management of construction enterprises under the new economic situation. *Financ. Econ.* **18**, 93–95 (2023)
9. Junquan, L.: Research on the relationship between tax management and economic benefits of construction enterprises. *Brand Res.* **26**, 185–187 (2023). <https://doi.org/10.3969/j.issn.1671-1009.2023.26.062>
10. Yanping, L.: Countermeasures analysis on strengthening the business administration level of construction enterprises under the new economic environment. *Caixun* **16**, 58–60 (2023)
11. Junmo, Z.: Correlation analysis of cost budget management and economic benefits in construction enterprises. *Economist* **5**, 85–86 (2022)
12. Valentyna, Y., Roman, Y.: System of organizational and economic support of human resources management at enterprises. *J. Vasyl Stefanyk Precarpathian Nat. Univ.* **6**(3–4), 112–120 (2019)
13. Zhi, L.: Discussion on the improvement measures of the economic management system of construction enterprises. *Build. Decorat.* **9**, 94–96 (2022)
14. Min, Y.: Path considerations for strengthening financial management in construction enterprises - based on the perspective of the new economic environment. *Caixun* **27**, 179–182 (2022)
15. Shuchen, L.: Discussion on the current situation and countermeasures of economic management in construction enterprises. *China Market* **7**, 78–79 (2022)



The Impact of Fintech Development on Commercial Banks' Credit Risk: The Case of SPD Bank

Xu Ruoyu, Wang Heng, and Xu Yanni^(✉)

Hubei Business College, Hu'bei, China

Xuyanni15@163.com

Abstract. Intech is showing strong momentum in 2024, with the market size continuing to expand, and technological innovations such as big data and artificial intelligence driving the continuous innovation of financial products and services, which are accelerating towards intelligence, efficiency, and universality, despite challenges such as data security.

Based on the above background, this paper, based on the research results at home and abroad, takes SPD Bank as an example, collects and organizes the annual report data of SPD Bank in the past ten years, and analyzes the association between the development of fintech and the credit risk of commercial banks through the empirical test, as well as analyzes the current fintech credit risk management of SPD Bank. Based on the theoretical analysis of the impact of fintech on the credit risk-taking behavior of commercial banks, the impact of the development of fintech on the credit risk of commercial banks is argued.

Keywords: Fintech · Big Data · Credit Risk · Commercial Banks · Non-performing Loan Ratio

1 Introduction

In recent years, the Chinese Government has taken a positive attitude towards the development of fintech and has introduced a number of policies to promote its healthy development. Fintech has played an important role in commercial banks' credit risk management, improving credit assessment accuracy and risk warning ability through big data and artificial intelligence technology, but also bringing new challenges such as data security and privacy protection. [3] Taking SPD Bank as an example, this paper discusses the multifaceted impact of fintech on commercial banks' credit risk, aiming to provide the industry with strategies to effectively address the challenges.

The study aims to analyze how FinTech affects credit risk management of commercial banks, enhances their ability to identify, assess and respond to risks, and strengthens their market competitiveness [1]. Fintech not only revolutionizes the banking business model and intensifies competition within the industry, but also brings many benefits to banks by building high-quality service platforms, optimizing customer experience and improving credit approval efficiency [2]. Taking SPD Bank as an example, this paper summarizes the

practices and innovations by analyzing and organizing the annual report data from 2012–2022, revealing the key role of FinTech in risk management, and providing commercial banks with experiences that can be drawn upon.

2 Fintech Development and Its Impact on Commercial Bank Credit Risk

2.1 Current Status of Financial Technology Development

Financial technology (Fintech), as an outcome of the profound integration between finance and technology, has witnessed remarkable advancements on a global scale in recent years. This fusion has given rise to innovative solutions that are transforming the financial landscape. Ba Shusong, Bai Haifeng (2016) summarized Fintech as technological innovations such as big data, cloud computing, artificial intelligence, blockchain, etc., and applying them to six major financial fields such as payment and settlement, asset management, commercial insurance, etc. [17]. Yang Wang and Qu Shuangshi (2017) similarly believe that blockchain, as the most cutting-edge technology of fintech, will provide new solutions for commercial banks to reduce information asymmetry and bring great changes to the financial ecology in the future [16]. Hong Juan (2014) analyzed big data and cloud computing technology and argued that commercial banks should build databases and use big data and cloud computing to process user information and transaction data to provide technical support for banks' micro-credit business [22].

Fintech in 2024 shows strong momentum of development, the market scale continues to expand, the first quarter of the global investment and financing activities are frequent, the total amount of \$ 17.5 billion, technological innovations such as big data, artificial intelligence, etc. to drive the continuous innovation of financial products and services, at the same time, the optimization of the policy environment and the improvement of the regulatory system for the development of the industry to protect, despite the challenges of data security, but the fintech is accelerating to the Despite the challenges of data security, Fintech is accelerating its development in the direction of intelligence, efficiency and universalization.

2.2 Theoretical Analysis of the Impact of Fintech Development on Commercial Bank Credit Risk

Fintech has significantly improved the credit risk warning and control capabilities of commercial banks through network vision and technological innovation [4]. However, its introduction has also brought about more complex risk challenges [5]. This is mainly reflected in the following three points:

2.2.1 Profitability and Risk Appetite

Fintech affects banks' profitability levels and willingness to take risks by optimizing business processes and innovating products [6]. When profitability improves, banks' risk tolerance decreases; conversely, in pursuit of profits, they may relax credit standards and increase credit risk [11].

2.2.2 Cost Control and Risk Tolerance

Fintech optimizes banks' ability to control costs. When costs are reduced, the bank's risk tolerance increases; if costs are poorly controlled, it may adopt a more aggressive risk strategy [7]. This mechanism works indirectly on banks' risk appetite by affecting their revenue and cost structure [10].

2.2.3 Risk Control Capacity

Fintech has revolutionized traditional risk control models and improved the efficiency of risk control systems, enabling banks to identify and manage risks more accurately [8]. This mechanism serves as a balance between performance pressure and risk control, ensuring that banks effectively manage credit risk while pursuing profits [9].

3 Model Analysis of the Impact of Fintech on the Credit Risk of China's Commercial Banks

3.1 Variable Setting

In this paper, the impact of FinTech development on credit risk is selected from the perspective of three mechanisms of action. The data of each variable are shown in Table 1 [13].

Explained Variable: non-performing loan ratio (NPL) is one of the key indicators for assessing the safety of banks' credit assets [14]. The higher its value, the greater the proportion of loans that cannot be recovered on schedule in the total loans. Given that bank NPLs are regarded as the main risk in China's financial sector, it becomes an important reference for measuring bank credit risk. Therefore, in our study, we choose the NPL ratio as a key indicator referring to the credit risk index.

Explanatory Variables: fintech index (FIN). In this paper, the Financial Inclusion Index constructed by Peking University is selected as the FinTech index, which usually includes digital channels, digital products, digital marketing, digital services and other aspects. Through this index, banks can assess their own digital development level. [18].

Control Variables:

Profitability indicator (ROA), numerous studies tend to use ROA or ROE to assess the earnings profile of financial institutions. In view of banks' unique liability business model, we choose ROA as a proxy to reflect banks' profitability [15]. Generally speaking, a higher level of ROA for a bank indicates better utilization of its assets, more significant profitability, and thus a stronger resistance to credit risk. The level of credit risk borne by commercial banks is affected by ROA to a certain extent, therefore, we comprehensively considered the importance of ROA and used it as a control variable in our study [19].

Loan Growth Rate (LGR) is an important indicator of the rate of expansion of a bank's business [20]. An increase in loan growth rate implies expansion of bank business and increase in income, which helps to reduce credit risk. However, too high loan growth rate may damage bank liquidity and lead to the rise of NPL, which in turn increases the credit risk of commercial banks. Therefore, loan growth rate has a dual impact on the credit risk of commercial banks.

Table 1. Data for each variable

particular year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
non-performing loan ratio	0.58	0.74	1.06	1.56	1.89	2.14	1.92	2.05	1.73	1.61	1.52
Shanghai Digital Inclusive Finance Index	150.77	222.14	239.53	278.11	282.22	336.65	377.73	410.28	431.93	458.97	460.69
Profitability indicators (%)	1.09	1.12	1.13	1.01	0.92	0.90	0.90	0.85	0.74	0.66	0.60
Loan growth rate (%)	16.01	14.43	14.76	10.70	23.04	15.63	11.10	11.91	14.15	5.56	2.39

Source: SPD Bank Annual Report, Peking University Digital Finance Research Center

3.2 Modeling

To ascertain the influence of SPD Bank's fintech evolution on its credit risk exposure, this study constructs a regression analysis model utilizing the bank's non-performing loan (NPL) ratio, profitability index, loan growth rate, and the Shanghai Digital Financial Inclusion Index spanning the period from 2012 to 2021. Specifically, the explanatory variable NPL represents the non-performing loan ratio, serving as a metric for assessing the bank's credit risk level. Additionally, the explanatory variable FIN captures the extent of the bank's fintech development. The remaining explanatory variables are characteristic variables of the bank, while ε denotes the random disturbance terms. The model is formulated as follows:

$$NPL = \beta_0 + \beta_1 fin + \beta_2 ROA + \beta_3 LGR + \mu \quad (1)$$

3.3 Model Test Analysis

3.3.1 Descriptive Statistics

Descriptive statistics for each variable are shown in Table 2:

3.3.2 Multicollinearity Test

Before using linear regression can be verified by statistical class methods such as normality test on the data to prevent pseudo-regression. In this paper, the data are tested for normality as well as the VIF test method, and the results are shown in the following Tables 3 and 4:

Table 3 test results shows that the NPL ratio (%) sample, ROA (%) sample, loan growth rate sample, and digital financial inclusion index sample are all $N < 5000$, so using the S-W test, the significance p-values are $NPL = 0.202$, $ROA = 0.394$, $LGR = 0.578$, and $FIN = 0.518$, respectively, and the levels do not show significance, so the the original hypothesis is not rejected and the data satisfy normal distribution.

The results of the tests in Tables 4 shows that the VIFs of the variables are all less than 10, indicating that there is no serious multicollinearity in the model.

Table 2. Descriptive statistics of variables

Variable name	Sample size	Maximum values	Minimum value	Average value	(Statistics) standard deviation	Upper quartile	Variance (statistics)
Digital Inclusive Finance Index	11	460.691	150.77	331.73	104.764	336.651	10975.478
Loan growth rate	11	0.23	0.024	0.127	0.055	0.141	0.003
ROA (%)	11	0.012	0.006	0.01	0.002	0.009	0
Non-performing loan ratio (%)	11	2.14	0.58	1.527	0.521	1.61	0.272

Table 3. Results of normality test

Variable name	Skewness	Kurtosis	S-W test
Non-performing loan ratio (%)	−0.819	−0.458	0.903 (0.202)
ROA (%)	−0.243	−1.063	0.928 (0.394)
Loan growth rate	−0.214	1.034	0.945 (0.578)
Digital Inclusive Finance Index	−0.267	−1.145	0.94 (0.518)

Table 4. Results of multiple covariance tests

Variant	ROA (%)	Loan growth rate	Data Inclusion Index
VIF	8.689	1.665	8.871

3.3.3 Heteroskedasticity and Autocorrelation Tests

The sample size is 11 and the model is a multiple linear regression model, which is tested for heteroskedasticity using White’s test. The results are shown in the following Table 5:

Table 5. White’s test results

F-statistic	2.505864	Prob. F (9,1)	0.4567
Obs*R-squared	10.53296	Prob. Chi-Square (9)	0.3091
Scaled explained Ss	2.450032	Prob. Chi-Square (9)	0.9822

It can be obtained that $nR^2 = 10.53296$, at $\alpha = 0.025$, check the distribution table of X^2 and get the critical value $X_{0.052}(9) = 19.02280 > nR^2$, indicating that the residuals are homoskedastic and there is no heteroskedasticity.

And the autocorrelation test of the model can be obtained as follows Table 6:

Table 6. LM test results

F-statistic	0.375681	Prob. F (2,5)	0.7047
Obs*R-squared	1.437049	Prob. Chi-Square (2)	0.4875

The observation of the p-value to $P = 0.4857 > 0.05$ indicates that the model is not autocorrelated.

3.4 Empirical Findings

After all kinds of tests on the regression model to get the model does not have serious multicollinearity, there is no heteroscedasticity and autocorrelation, so the model linear regression results are as follows Table 7:

Table 7. Results of linear regression analysis

	B	Standard Error	Beta	T	p
a constant (math.)	-2.190	2.365	-	-0.926	0.385
ROA (%)	76.498	148.453	0.303	0.515	0.622
Loan growth rate	5.899	2.454	0.620	2.404	0.047*
Digital Inclusive Finance Index	0.007	0.003	1.357	2.280	0.057
R^2	0.721				
Adjustment R^2	0.601				
F	$F(3,7) = 6.019$, $p = 0.024$				
D-W value	1.437				
Dependent variable: NPL ratio (%)					

Observing the model's goodness-of-fit, the R^2 value of 0.721 suggests a strong fit, indicating that the model effectively explains the variations in the dependent variable. Furthermore, the F-statistic of 6.019, significant at the 5% level, underscores the overall statistical significance of the model, affirming its robust construction. Consequently, the derived equation of the model is presented as follows:

$$NPL = -2.19 + 76.498ROA + 5.899LGR + 0.007FIN \quad (2)$$

Analyzing the regression outcomes presented in Table 7, the coefficient associated with FIN (0.007) attains statistical significance at the 1% level, signifying a notable impact of financial technology development on Pudong Bank's credit risk. Among the control variables, the coefficient of loan growth rate (5.899) is significant at the 5% level, indicating that an increase in loan growth exerts a significantly positive influence on credit risk. Similarly, the coefficient of ROA (76.498), also significant at the 5% level, demonstrates that a higher level of profitability has a markedly positive effect on credit risk.

4 Conclusions and Recommendations

4.1 Conclusion

First, the development of fintech has significantly increased the credit risk of commercial banks, prompting them to adopt more proactive and forward-looking strategies in risk management. Combined with the mechanism of cost control ability, it can be seen that the current increase in PBoC's credit risk is due to the fact that the initial stage of FinTech development requires banks to invest a large amount of capital, which increases the cost burden (Fig. 1). Under the business objective of pursuing profits, the bank may take on higher credit risk to maintain or improve profitability, leading to an increase in the level of credit risk. However, with the continuous development and maturity of fintech, banks' investment in this area will gradually level off and costs will decrease accordingly. At this point, banks' risk-taking behavior will also be adjusted and tend to be more robust. Therefore, the impact of fintech on bank credit risk is a dynamic process that requires comprehensive consideration of the characteristics and factors at different stages.

The limitations of this study are multifaceted. Firstly, the assessment of fintech's impact on commercial banks' credit risk is confined to short-term effects, neglecting the potential long-term implications. Secondly, given the disparities in internal credit risk management mechanisms among banks, the analysis lacks comprehensiveness in examining how fintech development differentially impacts various banks' credit risk profiles. Lastly, while SPD Bank, as a prominent domestic institution, offers valuable insights, the limited geographic scope of its operations limits the representativeness of the sample data. This could undermine the generalizability and reliability of the findings, particularly in large-scale data analyses, where ensuring a comprehensive sample becomes a crucial yet challenging aspect. Becomes a problem that cannot be ignored.

4.2 Recommendations

Amidst the rapid proliferation of fintech, it is imperative for governments to assume a proactive regulatory stance, particularly towards undercapitalized banks. By imposing restrictions on their fintech investments, the government can mitigate the risk of capital constraints hindering robust auditing standards and inadvertently intensifying credit risks [12]. Concurrently, the authorities should endeavor to foster the development of regulatory technology (RegTech), enabling swift and effective oversight of fintech-related hazards. This dual-pronged approach ensures that while embracing the convenience

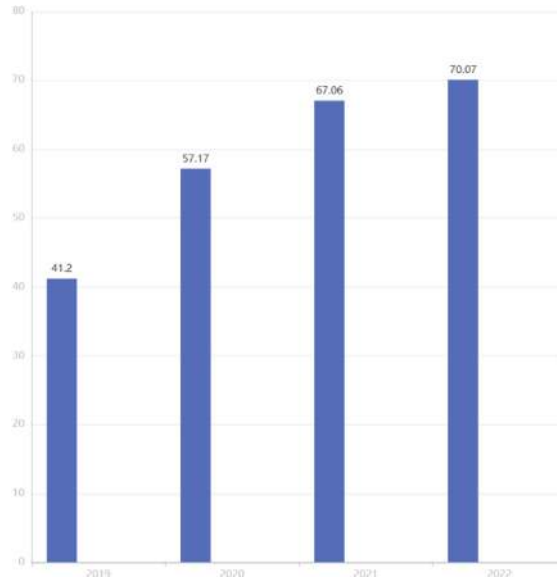


Fig. 1. Amount of FinTech Investment by POSCO, 2019–2022 (\$ billion)

and efficiency offered by fintech, the stability and security of the financial market are safeguarded, thereby fostering the healthy functioning of the entire financial system.

It is necessary to establish a scientific training mechanism for composite talents and a talent incentive mechanism to attract fintech talents to join with a highly competitive compensation system and a transparent promotion route. Commercial banks are in greater need of composite talents who can skillfully utilize fintech on the basis of mastering financial knowledge and familiarizing themselves with business processes [21]. After entering the era of fintech, the structure of financial products and service forms have become more and more complicated, and ordinary talents can hardly meet the requirements of listed commercial banks. By continuously introducing and cultivating fintech talents, commercial banks can more fully utilize fintech to empower their business activities and enhance their credit risk management capabilities.

References

1. Yudaruddin, R., Soedarmono, W., Adi Nugroho, B., et al.: Financial technology and bank stability in an emerging market economy. *Heliyon* **9**(5), e16183 (2023)
2. Hughes Joseph, P., Julapa, J., Moon Choon, G.: Consumer lending efficiency: commercial banks versus a fintech lender. *Financ. Innov.* (2022)
3. Maoyong, C., Qu, Y.: Does bank FinTech reduce credit risk? Evidence from China. *Pacific-Basin Finance J.* (2020)
4. Puschmann, T.: The rise of customer-oriented banking – electronic markets are paving the way for change in the financial industry. Rainer Alt. *Electronic Markets* (2012)
5. Yang, F., Hong, K.: Research on the impact of financial technology on commercial bank credit risk and its mechanism. *Financ. Dev. Res.* **06**, 66–73 (2022). <https://doi.org/10.19647/j.cnki.37-1462/f.2022.06.009>

6. Liu, L., Li, M.: Fintech and bank credit risk-based on empirical evidence from commercial banks in China. *J. Harbin Comm. Univ. (Soc. Sci. Ed.)* **02**, 26–42 (2022)
7. Zhu, X., Li, X.: Fintech and bank credit risk: exacerbation or reduction. *J. Shanxi Univ. Financ. Econ.* **44**(11), 39–52 (2022)
8. Tang, J., Li, H.: Research on the impact of financial technology on commercial bank credit risk. *Foreign trade Econ.* (12), 83–85+116 (2022)
9. Yao, T., Song, L.: Research on the economic capital impact of financial technology on commercial bank credit risk. *Sci. Technol. Manage. Res.* **41**(01), 104–110 (2021)
10. Biyun, R., Zongjie, Z.: The impact of financial technology on commercial banks' risk taking-based on the perspective of commercial banks' credit structure. *J. Guizhou Univ. Financ. Econ.* **05**, 61–69 (2021)
11. Yang, W., Zhu, S., Kwong, Y.: Fintech development, market competition and bank risk taking. *Financ. Theory Pract.* (03), 52–57 (2020)
12. Li, X.: Analysis of credit risk influencing factors and governance strategy of commercial banks. *Nat. Circul. Econ.* **23**, 147–149 (2020)
13. Ma B., Fan, R.: Impact of leverage regulation on credit risk of listed commercial banks in China--System GMM estimation based on dynamic panel model. *Econ. Issues* (01) (2019)
14. Cheng, C.: Research on the impact of financial technology on the risk-taking behavior of commercial banks in China. Wuhan University (2019)
15. Wang, K., Wu, Q., Li, K.: Fintech and commercial bank risk taking-an empirical analysis based on the Chinese banking industry. *Manage. Modernization* **37**(06), 100–104 (2017)
16. Gong, X., Yang, W., Qu, S.: Principles of blockchain technology and its application in the financial field, In: *International Monetary Review 2017 Combined Series*. Central Huijin Investment Limited Liability Company; Shenzhen Qianhai Hands Financial Technology Research Institute; Institute of International Monetary Research, Renmin University of China 15 (2017)
17. Shusong, B., Haifeng, B.: The development history of fintech and exploration of core technology application scenarios. *Tsinghua Financ. Rev.* **11**, 99–103 (2016)
18. Pin, G., Yue, S.: The impact of Internet finance on commercial banks' risk-taking: a theoretical interpretation and empirical test. *Financ. Trade Econ.* (2015)
19. Renmiao, Y.: Research on the impact of internet finance on credit risk management of commercial banks in China. *Financ. Econ.* **08**, 54–55 (2017). <https://doi.org/10.14057/j.cnki.cn43-1156/f.2017.08.024>
20. Wu, S., Zhu, Y., Li, T.: Interest Rate Marketization, Internet Finance and Commercial Bank Risk - An Empirical Test Based on Panel Data Dynamic GMM Method. *Res, Financ. Econ.* **30**(06), 29-38 (2015)
21. Wang, K.: Fintech, price competition and bank risk taking. *J. Harbin Univ. Comm. (Soc. Sci. Edn.)* **01**, 40–48 (2018)
22. Hong, J.: Research on competitive strategies of commercial banks' "micro and small loans" under big data environment. *Southern Financ.* (03), 80–82+9 (2014)



Exploring the Avenues for Optimizing University Education Management Through the Empowerment of Information Technology

Shuqing Wang^(✉)

Guangdong University of Science and Technology, Guangzhou, China
394499948@qq.com

Abstract. The development of informatization has had a significant impact on the education industry, and many universities have begun to use informatization methods to carry out management work. To explore the optimization path of information technology for university management, this paper constructs a data-driven informatization evaluation system for university education management and discusses the optimization methods and approaches for informatized university education management based on this.

Keywords: Informatization · Technology · University Management · Optimization Path

1 Introduction

With the progress and development of the times, various industries are entering the informatization stage. Educational informatization is an important way to achieve the modernization of education in China, and it is also a strategic choice in the development of China's education industry [1]. As an essential part of educational informatization, the informatization of university education management is a component that requires continuous innovation and optimization. In today's era of continuous internet development, universities using informatized management will inevitably generate a large amount of data. These data need to be scientifically analyzed and mined to reflect their value, which is crucial for exploring the optimization path of university education management.

Currently, many universities in China have implemented informatized management, but the level of informatization varies [2]. Without a specific informatization evaluation plan, the informatization level of university education management cannot be objectively and accurately evaluated, and the optimization path cannot be explored. This paper utilizes the concept of data-driven through the analysis of relevant literature to construct a reasonable evaluation system for the current informatization of university education management. Based on this, we explore the optimization path of informatization in university education management.

2 Research Background

Against the backdrop of rapid information technology development, countries around the world have embarked on plans to apply informatization to the education industry [3] and incorporated this initiative into their national strategic objectives. Internationally, the United States invested \$200 million in the “Big Data Research and Development Initiative” in 2012. In 2014, the United Kingdom issued a document stating that data-driven development will become an essential force for reform in the new era. Similarly, in 2014, Japan’s “Talent Cultivation in the Big Data Era” proposed the establishment of specialized educational institutions that focus on data. Domestically, the “Action Plan for Promoting Big Data Development” issued in 2015 clearly instructed the implementation of a big data project for education and culture. The concept of “driving educational modernization through educational informatization” was also put forward in 2001. Therefore, both domestically and internationally, the informatization of educational information management has always been a focal point of attention in education for various countries.

Currently, many experts and scholars have conducted research on the informatization evaluation of university education management [4], as most scholars in the field agree that accurate and reasonable evaluation is an important way to promote the optimization and progress of educational management informatization. However, the informatization evaluation systems used in most universities are still relatively simple, and the evaluation system is not complete enough. The value of process data has not been effectively explored, which is not conducive to the future development of informatization in university education management.

3 Related Concepts and Technologies

3.1 Data-Driven

Data-driven refers to the approach that takes lean analysis and the concept of a closed-loop data system as its foundation. After business data is digitalized, the collected data is treated as a production material. Through data analysis and mining, the intrinsic value of the data is extracted and then applied to practical business operations, promoting business optimization and realizing data-centric business development. Both lean analysis and the closed-loop data system are concepts where data serves as a production factor. Data collection, data analysis, and data mining are the methods and means of data-driven approaches, with data-centric decision-making and actions being the goals of data-driven efforts [5].

3.2 Introduction to Hadoop

Hadoop is a distributed infrastructure developed by the Apache Foundation in 2002. Hadoop can process large amounts of data in a distributed manner, which is relatively efficient compared to other methods. This is because all data is processed in parallel at high speed, and multiple maintained copies of the processed data ensure high reliability.

Due to Hadoop's scalable nature, it can handle petabyte-scale data [6]. Additionally, despite its many advantages, Hadoop has a low cost of use, making it a popular choice for individuals and businesses alike. The main components of Hadoop include HDFS (Hadoop Distributed File System), Hive (a data warehouse tool), MapReduce (a parallel computing model), Sqoop (an open-source tool), and Flume (a log collection system). The specific ecosystem architecture diagram of Hadoop is shown in Fig. 1.

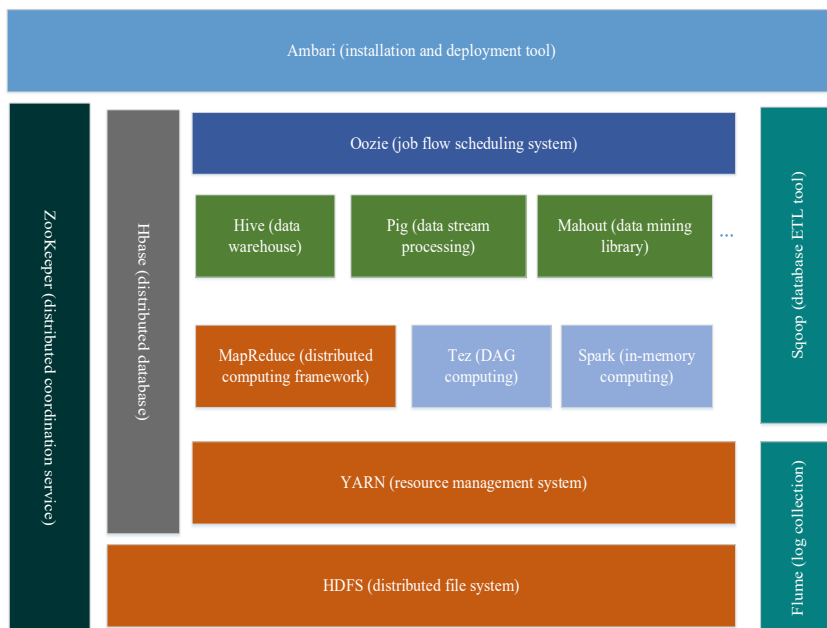


Fig. 1. Hadoop Ecosystem Architecture Diagram

3.3 Distributed File System HDFS

The Hadoop Distributed File System (HDFS) serves as the underlying structure of Hadoop, responsible for storing files across all nodes. Due to its high fault tolerance, HDFS is ideal for applications involving large-scale datasets [7]. HDFS follows a master-slave architectural model, where a NameNode acts as the master server, managing the file system's namespace and handling client file access operations. Multiple DataNodes manage the stored data. Additionally, a Secondary NameNode assists the NameNode by sharing its workload. The specific structure is illustrated in Fig. 2.

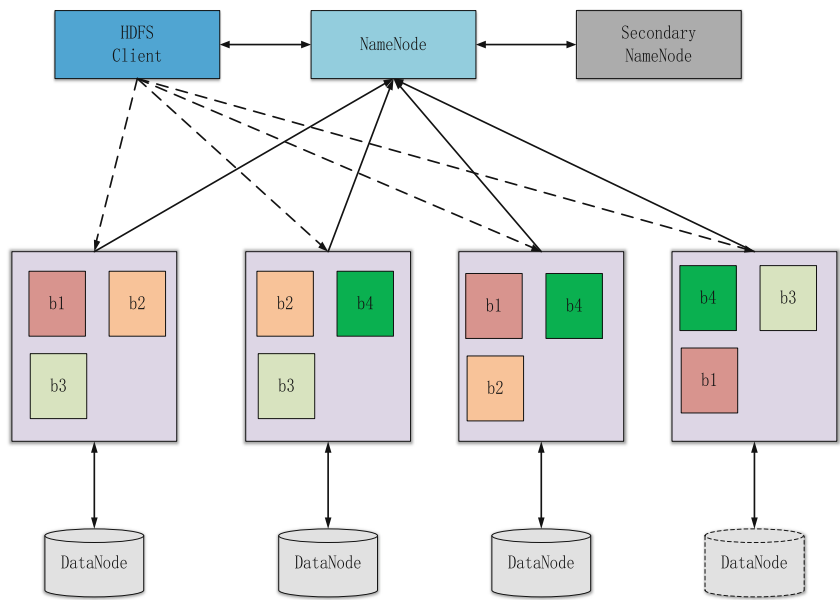


Fig. 2. HDFS Architecture Diagram

4 Construction of an Evaluation System for the Informatization Level of Education Management in Colleges and Universities

4.1 Construction Principles

4.1.1 Scientificity

The principle of scientificity in system construction means that the establishment of an evaluation index system should not be without basis and should be constructed according to scientific methods. In the process of construction, it is necessary to follow the objective development laws of informatization, take into account the current situation and characteristics of informatization in university management, identify future development directions from international and domestic informatization research, and enhance the scientific attribute of the system.

4.1.2 Comprehensiveness

When planning the construction, comprehensive consideration should be given to all aspects. Every link and process of the system should be carefully planned, and all elements in the research process should be reflected, while avoiding redundancy. The various elements should be interconnected to improve the efficiency of informatization. The first-level indicators for each business in the evaluation system should be determined, and then the details below should be improved based on these first-level indicators.

4.1.3 Systematic Principle

The entire system should be treated as a whole, where each indicator cannot exist in isolation [8]. There should be connections and influences among them, but not overly so, maintaining a relatively independent state. Based on such requirements, a hierarchical system should be established with clear relationships between levels, carefully classifying each level of indicators from high to low, and from partial to global perspectives.

4.1.4 Operability

The constructed system should not only exist theoretically but also be practical and operable. If there is data that is difficult to collect, it can be considered for removal [9]. Each indicator in the system should follow the principle of simplicity. Intuitive indicators can avoid interference from other factors during scientific quantification of data, resulting in clearer and more understandable evaluation results. Factors that improve system operability should be considered as much as possible during the construction of the indicator system.

4.1.5 Developmental Adaptability

When constructing the system, the user group and scope of the indicators should be considered. It should not only be usable within a small scope but should be adaptable to all research subjects. Moreover, the system should not have excessively high requirements for the usage environment and should be widely applicable. Since informatization is constantly evolving dynamically, the system also needs to be continuously updated to adapt to the development trends of management informatization [10].

4.2 Construction Method

Currently, most evaluation indicators for information management systems in universities are based on outcome-based data, while there are no clear indicators for process-based data, and the value inherent in process-based data indicators has not been fully explored [7]. To address this issue, this study reviewed relevant domestic and international literature, considering both outcome-based and process-based data within the indicator system. By applying bibliometrics, pre-survey methods, and combined weighting methods, a data-driven evaluation system for university education management informatization was constructed. The specific process is illustrated in Fig. 3.

5 Design of a Data-Driven Evaluation System for the Informatization Level of University Education Management

5.1 Needs Analysis

The design goal of this system is to collect data from university education management, obtain reliable data through data analysis and cleaning, conduct evaluations, and finally present the evaluation results intuitively. In terms of performance, the system should be

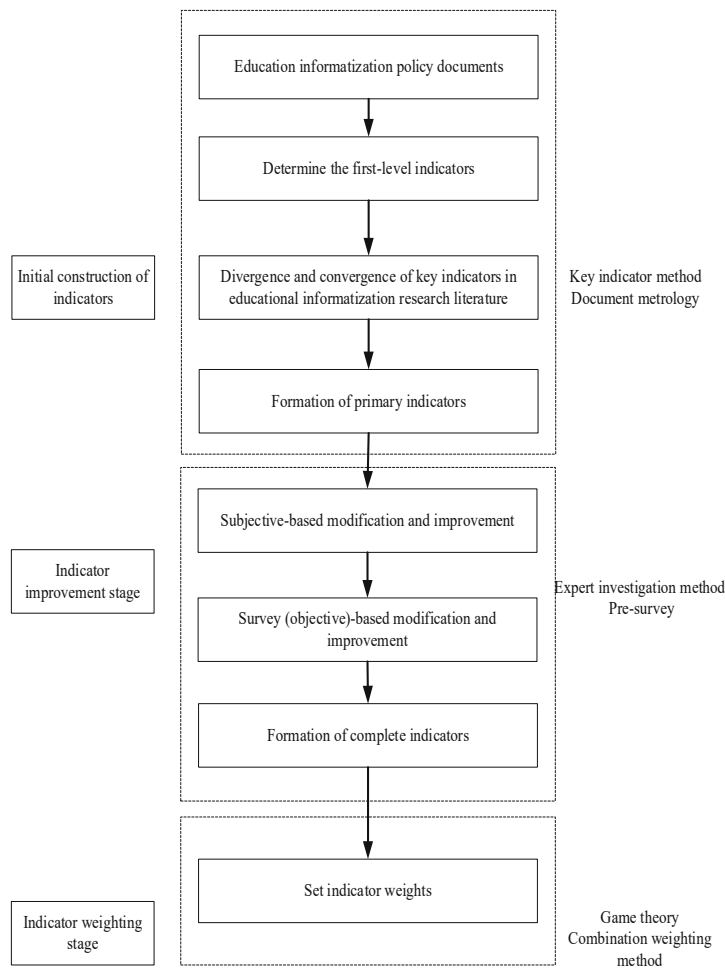


Fig. 3. Construction process of a data-driven evaluation index system for the informatization level of university education management

able to process data quickly, possess a certain degree of scalability, and handle bugs that may arise during system operation [11].

5.2 Overall Architecture Design

The evaluation system designed in this study mainly consists of four parts, as shown in Fig. 4.

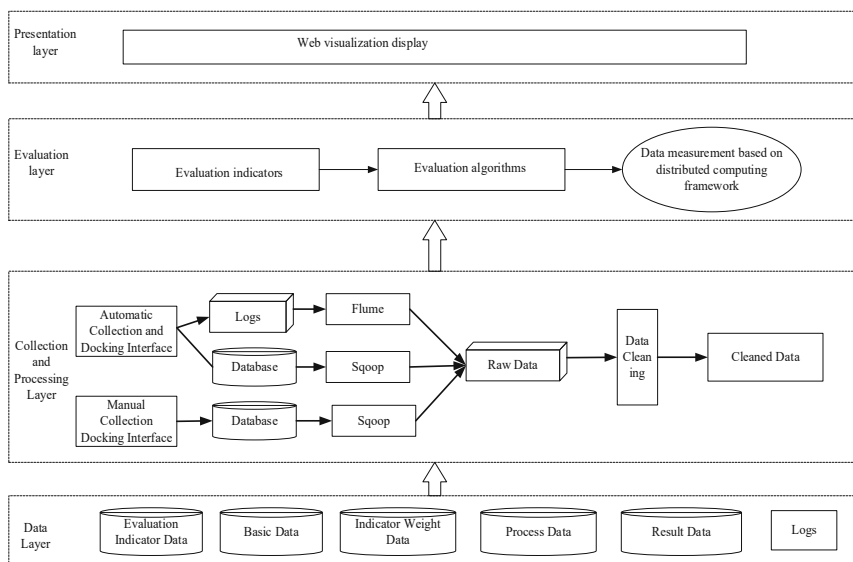


Fig. 4. Overall System Architecture

5.3 Data Acquisition Module Design

Data acquisition is the primary step, and there are two methods of collection: manual and automatic [12]. If there is data that cannot be collected automatically, manual collection will be used as a supplement. Manual collection primarily involves gathering data through questionnaire surveys, while automatic collection mainly utilizes Sqoop and Flume to capture data from databases and logs. See Figs. 5 and 6 for details.

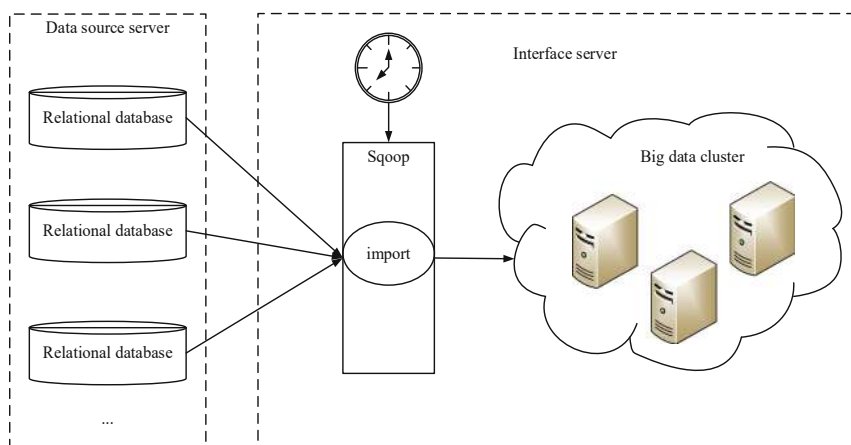


Fig. 5. Data Acquisition Approach with Sqoop

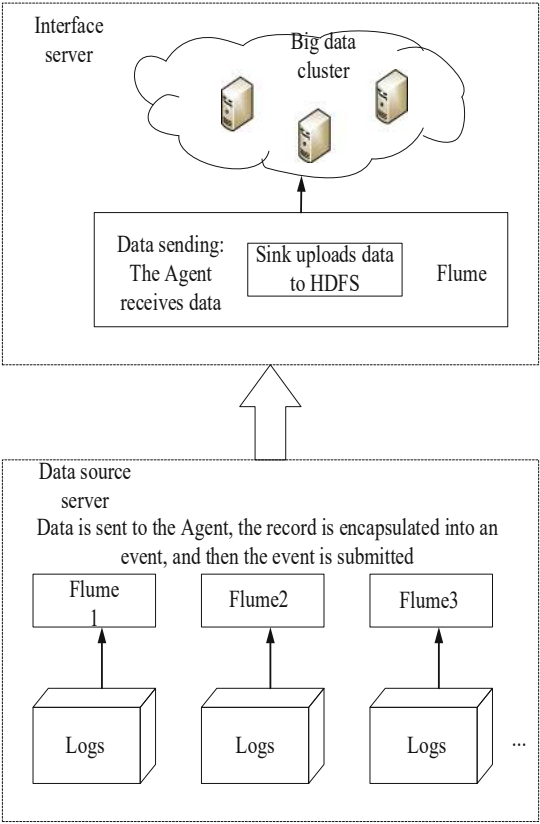


Fig. 6. Data Acquisition Approach with Flume

5.4 Data Cleaning Module Design

Data collected during the automatic acquisition process may contain duplicates or other issues [13], which can affect the accuracy of evaluations. Therefore, data cleaning is necessary. Given the large amount of data involved in the evaluation process, the Hadoop and MapReduce frameworks can be utilized for data processing. The specific approach is illustrated in Fig. 7.

5.5 Comprehensive Evaluation Module Design

The results of the comprehensive evaluation reflect the degree of informatization in university education management. The measurement data comes from the collected and cleaned data, mainly involving two modules:

- (1) Measurement of the development of tertiary indicators
The measurement of tertiary indicators refers to converting the obtained data into scores for tertiary indicators according to certain standards. Different indicator data corresponds to different calculation rules.

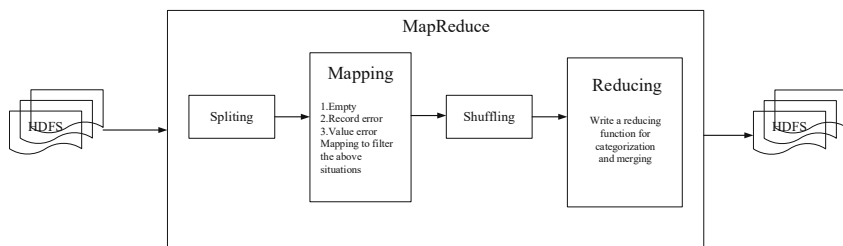


Fig. 7. Design Idea of Data Cleaning

(2) Comprehensive Evaluation

For the informatization evaluation of university management, this paper uses the comprehensive index method [14]. In the evaluation, the interrelationship between each indicator is not significant, and their importance is relatively equal. Therefore, the comprehensive index can be calculated using a linear weighted model. The calculation formula is shown in formula (1):

$$EDI = \sum_{i=1}^n W_i \left(\sum_{j=1}^m W_{ij} \left(\sum_{k=1}^r W_{ijk} Z_{ijk} \right) \right) \quad (1)$$

In the above formula, EDI represents the degree of informatization of university education management; n represents the number of first-level indicators in the level of informatization management; m represents the number of indicators included in the i th type of indicator in the level of management informatization; r represents the number of indicators included in the j th type of indicator in the level of management informatization, where j is the final-level indicator; W_i represents the weight of the i th type of indicator in the overall indicators, with $\sum_{i=1}^n W_i = 1$; W_{ij} represents the weight of the j th indicator in the i th type of indicator, with $\sum_{j=1}^m W_{ij} = 1$; and Z_{ijk} represents the dimensionless value of the k th indicator of the j th index in the i th category, with $\sum_{k=1}^r W_{ijk} = 1$ [15].

6 Application of Evaluation of Education Management Informatization Level

6.1 Evaluation Objects and Scheme

The evaluation object selected this time is the education management informatization system of a certain university, which includes a total of 40 management departments within the university. Based on the evaluation system studied above, the designed evaluation system is used for data collection and processing in all management departments. Finally, calculations are performed to understand the degree of education management informatization in the university through the measurement results. This allows for the analysis and study of optimization plans and approaches for the university's informatization education management. The specific evaluation process is shown in Fig. 8.

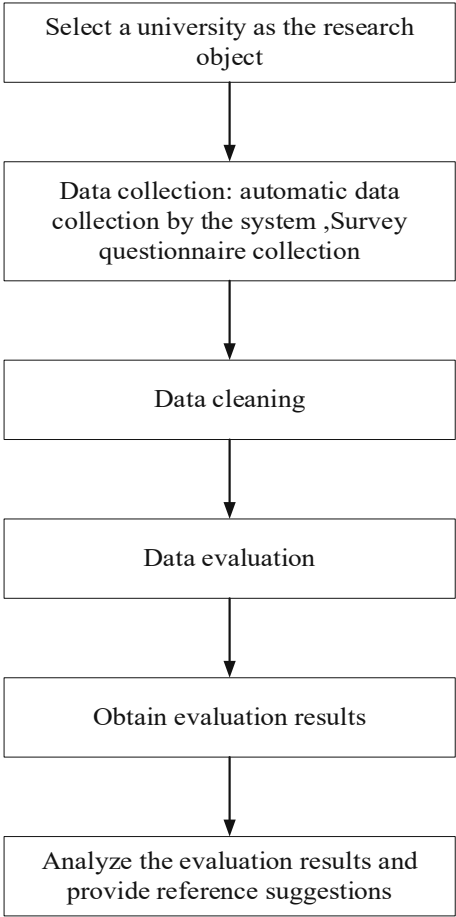


Fig. 8. Evaluation Process Design

6.2 Result Analysis and Suggestions

The evaluation results of information management in a university obtained from the evaluation system show that the average comprehensive index of informatization evaluation across various management departments in the university has reached 76.6. Moreover, 58.3% of the departments have evaluation results exceeding the average, indicating that more than half of the management departments have a relatively good level of management informatization. However, there are still over 40% of departments whose management informatization level has not reached the average. Based on these results, we can analyze and study the existing problems and optimization approaches in the informatization of education information management in the university. Firstly, the informatization level of various management departments within the university is uneven. It is recommended to intensify the popularization of business system construction, plan systems

according to the business needs of each department, and optimize and improve the supporting environment for informatization management. Secondly, the service informatization level of many departments is still relatively low. It is necessary to further integrate informatization technology into existing businesses, improve the efficiency of transaction and inquiry services, simplify work processes and steps as much as possible, and enhance the overall quality and efficiency of informatization management. Finally, it is essential to maintain management informatization, ensure the security of the network environment and data, and provide support for further optimization of informatization in university education information management.

7 Conclusion

This article provides research suggestions on optimizing the management of higher education through informatization technology, focusing on the establishment of an evaluation system. To explore the optimal paths for informatization management, it is crucial to first understand the current state and level of management informatization in universities. By analyzing the assessment results, we can identify existing problems and design targeted solutions based on those issues. This approach ensures effective optimization of informatization education management in universities. It is hoped that this study can offer an efficient avenue for optimizing educational informatization management across various universities.

References

1. Dan, X., Zeyi, W.: Thoughts and exploration on the informatization construction of asset management in universities in the new era – a review of “asset management in universities under the background of ‘double first-class’ construction.” *Sci. Technol. Manage. Res.* **43**(5), I0003 (2023)
2. Bing, L., Xiao, Y.: Analysis of the informatization path of higher education management in the context of big data – A review of “Higher Education Management and Its Informatization Construction in the Era of Big Data.” *J. Chin. Soc. Educ.* **4**, I0013–I0013 (2023)
3. DATA INTERNATIONAL: The DAMA Guide to the Data Management Body of Knowledge, pp. 37–56. Technics Publications LLC, New Jersey (2010).
4. Airong, W.: Exploration of the informatization construction of university archives management in the Internet age – A review of “Research on Modern Archives Management and Preservation Strategies.” *Sci. Technol. Manage. Res.* **43**(11), I0003–I0003 (2023)
5. Shanshan, Y.: Innovative path of college student management from the perspective of informatization – a review of “higher education informatization management and student management work.” *China Sci. Technol. Paper* **17**(4), I0015 (2022)
6. Li, X.: Innovation and practice of higher education management mode in the information age – a review of “exploration of higher education management in the new era.” *People’s Yangtze River* **53**(9), I0007 (2022)
7. Juan, H.: Development ideas for the transformation of higher education management in the era of big data – A review of “Research on Higher Education Management and Innovation Practice.” *Leadership Sci.* **12**, 145 (2022)
8. Dawei, X.: Exploring Innovative Paths of Higher Education Management in the Context of Informatization. *Food Res Dev* **44**(20), I0014 (2023)

9. Dongyun, L.: Innovation of teaching management mode and system in Colleges and universities in the new era – a review of “innovation research on teaching management mode in colleges and universities.” *J. Chin. Soc. Educ.* **4**, I0035–I0035 (2023)
10. Sanfa, C., Xiaoqi, J., Yumeng, L.: Information innovation path of university research from the perspective of data governance – taking t university as an example. *Modern Educ. Manage.* (2023). <https://doi.org/10.16697/j.1674-5485.2023.01.003>
11. The Central Cybersecurity and Informatization Committee issued the. In: 14th Five-Year National Informatization Plan. *Chinese Digit. Med.* **17**(1), 118 (2022)
12. Lu, X.: Analysis on the application of business intelligence in decision-making of Chinese universities: from the perspective of evidence-based decision-making. *Chongqing Higher Educ. Res.* **2**, 66–76 (2019)
13. Weber, K., Otto, B., Österle, H.: One size does not fit all---a contingency approach to data governance. *J. Data Inform. Qual.* **1**(1), 1–27 (2009). <https://doi.org/10.1145/1515693.1515696>
14. Tongshan, C.: How to improve university decision-making performance: from the perspective of institutional research and “data-driven decision-making” model. *Fudan Educ. Forum* **11**(2), 54–60 (2013)
15. Xiaohui, D., Xiaobin, Z., Yiping, P.: Framework design and implementation of big data governance in higher education. *Chinese Audio-visual Educ.* **8**, 63–71 (2019)



Research on the Development and Innovation Strategies of E-commerce Platforms Driven by Big Data Technology

Ruodan Luo¹, Ke Chen², and Chao Wang³(✉)

¹ School of Management, Wuhan Donghu College, Wuhan, China

² School of Economics, Wuhan Donghu College, Wuhan, China

³ Wuhan University of Engineering Science, Wuhan, China

wc_gckjxy@163.com

Abstract. With the advent of the digital economy era, big data technology has become an important driving force in the field of e-commerce, profoundly affecting the development and innovation of e-commerce platforms. This paper analyzes the importance and role of big data technology in the development of e-commerce platforms, discusses the development strategies of e-commerce platforms supported by big data technology, including the basic principles and goals of e-commerce platform development, the architecture design of e-commerce platforms based on big data technology, and the functional modules and development process of e-commerce platforms. In order to promote the continuous development and innovation of e-commerce platforms driven by big data technology, several recommendations are proposed, including precision marketing and personalized recommendations, intelligent customer service and after-sales services, supply chain optimization and inventory management, data security and privacy protection. It is hoped that the results of this paper can provide reference and inspiration for practitioners and researchers in the e-commerce industry, and promote further development and innovation of e-commerce platforms in the era of big data.

Keywords: Big Data Technology · E-commerce Platform · Digital Economy · Data Analysis · Personalized Recommendations

1 Introduction

With the advent of the digital economy era, big data technology, as an important branch of information technology, is increasingly becoming a key driving force for enterprises to grasp core competitiveness and promote innovative development. In the field of e-commerce, the application of big data technology (hereinafter referred to as BDT) has penetrated into various aspects of e-commerce platforms, posing new requirements and challenges for the development and innovation of e-commerce platforms. Nowadays, with the continuous growth of data volume and the rapid progress of data processing technology, e-commerce platforms are actively utilizing BDT to explore the value of massive data. On the one hand, through the collection, storage, processing, and analysis

of data, e-commerce platforms can build more accurate user profiles, better grasp market trends and user needs, and provide personalized recommendation services for users. On the other hand, in-depth analysis of user behavior can help e-commerce platforms optimize website layout, product display, and marketing strategies, thereby further improving user experience and promoting the growth of transaction conversion rates. This user centered personalized customization and differentiated service not only enhances user engagement and loyalty, but also significantly enhances the market competitiveness of e-commerce platforms [1].

However, facing the opportunities and challenges brought by big data, how e-commerce platforms effectively integrate and utilize these data resources to promote continuous innovation and development has become a focus of attention in both industry and academia. In the digital economy era, BDT is not only a crucial support for enterprise informatization but also a key driver for enterprise innovative development [2]. E-commerce platforms, as important carriers and windows of the digital economy, must fully leverage the advantages of BDT to achieve their sustainable development and innovation. Therefore, this paper explores the development and innovation strategies of e-commerce platforms driven by BDT, discussing how to fully utilize the advantages of BDT to achieve the sustainable development and innovation of e-commerce platforms.

2 Overview of BDT

2.1 Definition and Characteristics of BDT

With the deepening development of the information age, BDT, as a new type of information processing and analysis technology, is gradually becoming an important tool for enterprises and organizations to obtain real-time data insights and improve business decision-making capabilities. BDT refers to the use of specific techniques to process data sets that are difficult to manage and process using conventional methods. These data sets typically have characteristics such as large volume, fast generation speed, diverse types, and low value density. They involve technical methods and tools for processing and analyzing massive and diverse data, including data collection, storage, processing, and analysis. Through efficient distributed computing, machine learning, and data mining technologies, valuable information and insights can be extracted to support enterprise decision-making. It mainly deals with massive data, which comes from various channels and business systems, including structured data, semi-structured data, and unstructured data. Traditional data processing methods cannot handle such a large amount of data. However, BDT can effectively process these massive amounts of data and explore its potential value. BDT, with its powerful processing capabilities, can address the challenges of data diversity. This diversity is not only reflected in the breadth of data sources, but also encompasses the diversity of data types and differences in data formats. BDT can flexibly handle various types of data and provide comprehensive insights into the market and users for enterprises through integration and analysis across data sources. More importantly, BDT has the ability to process real-time data streams, enabling enterprises to quickly respond to data changes, conduct real-time analysis, and provide more timely data support for enterprise decision-making. In addition, BDT not only focuses on the massive amount of data, but also places greater emphasis on the comprehensiveness

and deep mining of data. It can conduct data mining and analysis from different dimensions and perspectives, revealing the inherent connections and potential patterns between data, and providing enterprises with deeper market insights and decision-making basis. BDT has enormous advantages and potential in processing, analyzing, and utilizing massive amounts of data, bringing more opportunities and challenges to enterprises and organizations. In the field of e-commerce, e-commerce platforms can fully utilize the advantages of BDT to achieve powerful functions such as personalized recommendations, user behavior analysis, and market trend prediction, providing strong support for the development and innovation of e-commerce platforms [3].

2.2 Core Technologies and Application Areas of BDT

The core technology of BDT covers the entire process from data collection, storage, processing to analysis, mining, and visualization, forming a complete data processing ecosystem. The framework of BDT is shown in Fig. 1. In the data collection stage, BDT achieves effective integration of structured, semi-structured, and unstructured data through ETL tools and API interfaces. Whether it is the internal data warehouse of the enterprise, user comments on social media, or real-time data generated by IoT devices, they can all be captured and integrated by BDT. In terms of data storage, distributed storage technologies such as Hadoop HDFS and NoSQL databases provide reliable storage solutions for massive amounts of data. Data processing technology uses computational models such as SQL and MapReduce to clean, transform, and compute data to meet analysis needs. Finally, through the application of algorithms and models, BDT can discover the correlations and trends between data, providing scientific basis for enterprise decision-making.

2.3 Application Status of BDT in E-commerce Platforms

With the rapid development of Internet technology and the increasing trend towards digitization, the application of BDT in the field of e-commerce is continuously deepening and expanding. By collecting user browsing behavior, purchase records, and interest preferences through BDT, e-commerce platforms can utilize intelligent analysis and algorithm calculations to recommend personalized products and services to users [4]. For example, e-commerce platforms like JD.com and Taobao have implemented precise personalized recommendations through BDT, thereby enhancing user purchasing experience and transaction conversion rates. In terms of personalized services and recommendations, BDT acts as a thoughtful shopping assistant, accurately capturing users' interests and needs. Alibaba Group is a vivid example, successfully achieving personalized recommendations through in-depth analysis of user behavior using BDT, thereby optimizing product displays and marketing strategies. This not only enhances user experience but also brings higher conversion rates for enterprises.

In the field of transaction security, BDT plays the role of a guardian. By real-time monitoring and analyzing transaction data, BDT can quickly identify potential risks and fraudulent activities, ensuring the safety and reliability of every transaction. Payment platforms like Alipay rely on BDT to implement risk warnings and anti-fraud functions, providing users with a safer and more trusted transaction environment. BDT can also help

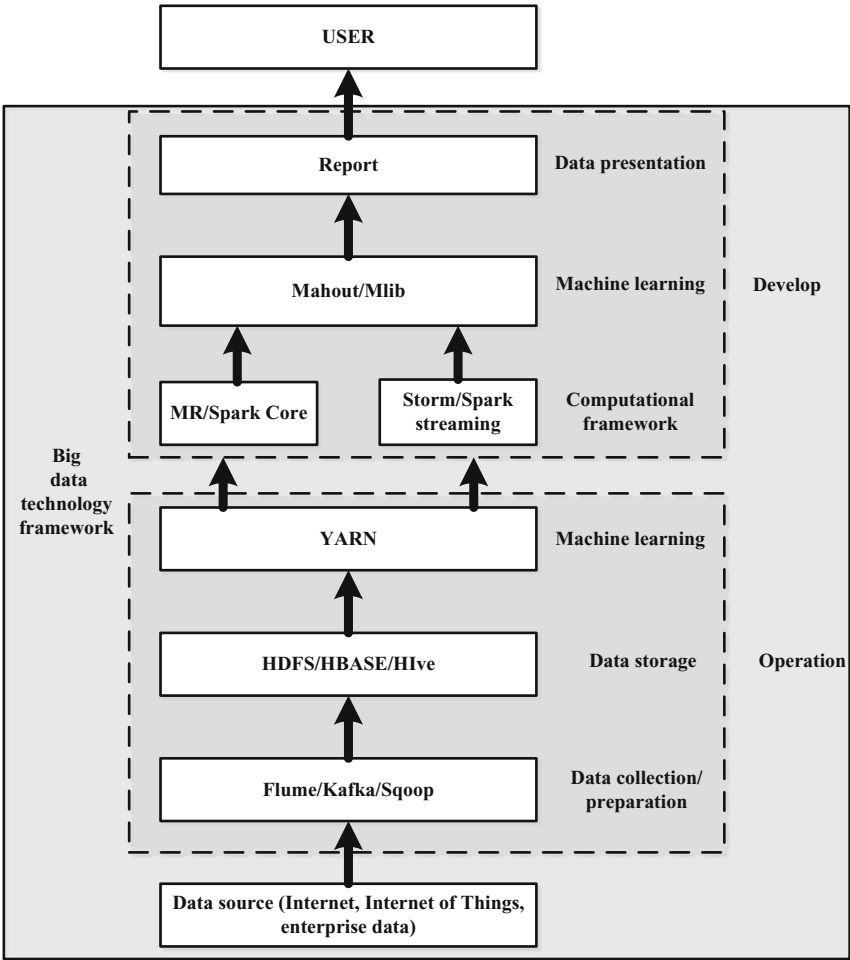


Fig. 1. BDT framework.

enterprises grasp the pulse of the market and predict future trends. Through deep mining and analysis of massive data, enterprises can insight into emerging market trends and consumer hotspots, thereby adjusting product strategies and marketing plans in a timely manner. E-commerce giants like Amazon rely on BDT to accurately predict market demands, achieving product positioning and optimization goals. Today, the application of BDT in e-commerce platforms has become an important means to enhance user experience, optimize marketing strategies, and improve transaction security.

3 Development Strategy for E-commerce Platforms Based on BDT

3.1 Basic Principles and Objectives of E-commerce Platform Development

For the development of e-commerce platforms, user experience should take precedence, and the development process should revolve around designing features and interfaces based on user needs to enhance satisfaction. Utilizing BDT for data-driven decision-making is essential to improve the platform's operational efficiency and profitability. It is also crucial to ensure the security of transaction data and protect privacy to prevent user information leakage. E-commerce platforms need to continuously innovate, introduce new features and services to maintain competitiveness. Regarding the objectives that should be achieved in developing e-commerce platforms, emphasis should be placed on enhancing the platform's personalized recommendation capabilities. By analyzing user behavior and preferences through BDT to achieve personalized recommendations, the purchase conversion rate can be increased. Improving interface design and feature settings based on data analysis and user feedback can enhance user retention rates. Additionally, real-time monitoring of transaction data using BDT to identify and handle risks enhances transaction security. Implementing intelligent marketing through market and user behavior analysis using BDT enables precise marketing strategies to increase sales and market share. Finally, continuous innovation is crucial by introducing new technologies and services to meet user demands and expand business scope [5]. Adopting a data-driven development model during the e-commerce platform development process, starting from data for requirements analysis and feature design, improves development efficiency and product quality. Prioritizing meeting user experience needs, continually enhancing design features through user research and feedback to increase user engagement and loyalty is important. Strengthening data security measures through encryption technology and secure validation to safeguard user transaction information is necessary. Continued innovation is also vital, staying informed about market trends and changes in user demands, adjusting strategies promptly, launching new products to maintain a leading position.

3.2 Architecture Design of E-commerce Platform Based on BDT

The architecture of an e-commerce platform based on BDT should include multiple aspects such as data collection and storage, data processing and analysis, platform service and application layer design, as well as security and monitoring mechanisms, as shown in Fig. 2. At the level of data collection and storage, e-commerce platforms accumulate a large amount of data during operation, such as user behavior records, detailed product information, etc. In order to effectively utilize this data, technologies such as distributed databases or cloud storage can be introduced, which not only ensure the security and reliability of the data, but also maintain efficient and stable operation when the data volume surges. In terms of data processing and analysis, by constructing appropriate data processing pipelines and analysis models, e-commerce platforms can achieve efficient processing and deep analysis of massive data. By using data processing techniques (such as Hadoop, Spark, etc.) and data analysis tools (such as machine learning algorithms, data mining techniques, etc.), real-time monitoring of user behavior, optimization of

recommendation algorithms, and prediction of sales trends can be achieved, thereby improving the operational efficiency and performance of the platform. In terms of platform services and application layer design, by carefully designing service interfaces and application modules, efficient collaboration between different functional modules can be achieved, improving the flexibility and maintainability of platform applications. With the help of microservice architecture or containerization technology, services can be quickly deployed and expanded to meet increasingly complex and diverse business needs [6]. In terms of security and monitoring mechanisms, e-commerce platforms can leverage BDT to achieve real-time monitoring of user behavior, identify potential security risks and respond in a timely manner, while strengthening data encryption and permission management to protect the security of user personal information and transaction data. Through reasonable design and implementation, a stable, efficient, and secure e-commerce platform can be built to improve platform operational efficiency, optimize user experience, and maximize commercial value, providing strong support for the rapid development of enterprises.

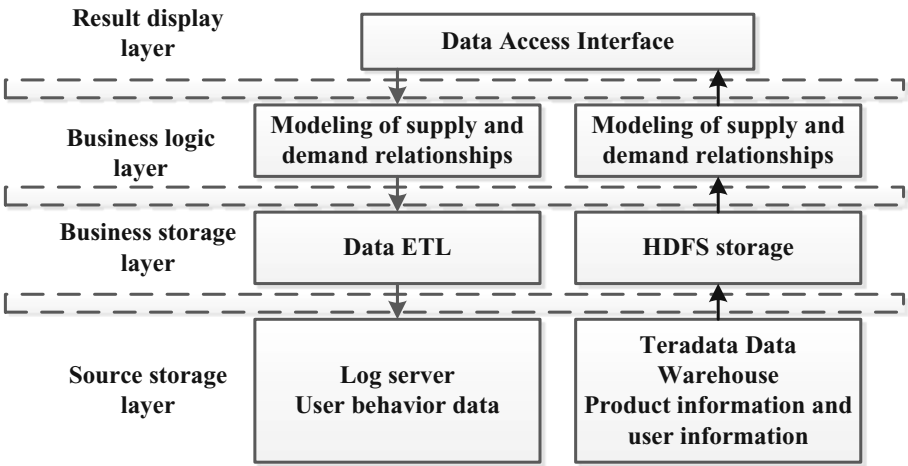


Fig. 2. Architecture design of e-commerce platform based on BDT.

3.3 Functional Modules and Development Process of E-commerce Platforms

E-commerce platforms generally include the following functional modules: user management module, product management module, shopping cart module, order management module, payment module, recommendation system module, data analysis module, etc. These functional modules are interrelated and together form a complete e-commerce platform system. The recommendation system module and data analysis module are two very important modules of the platform. The recommendation system module recommends personalized products based on user behavior data; The data analysis module analyzes user behavior and transaction data through BDT, providing business decision

support [7]. The functional modules of the B2C e-commerce platform are shown in Fig. 3.

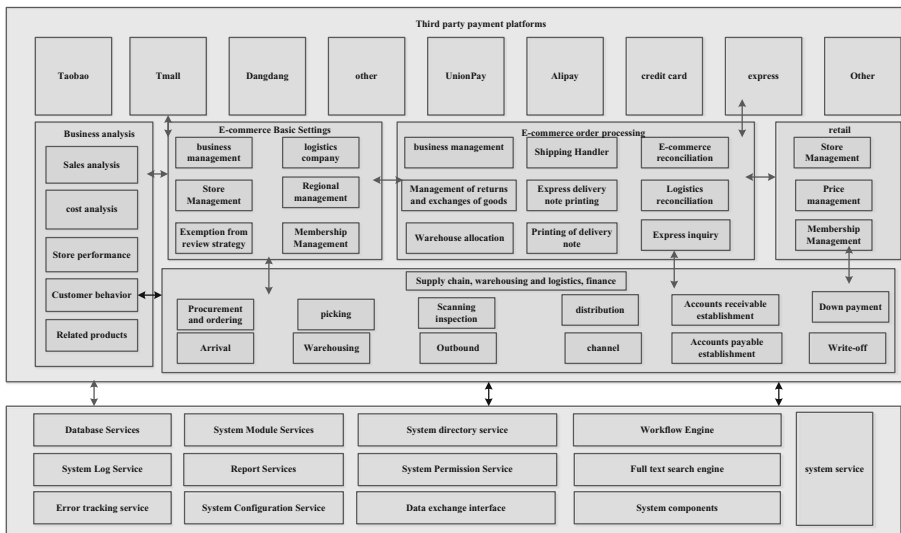


Fig. 3. Functional modules of B2C e-commerce platform.

The development process of e-commerce platforms usually includes requirements analysis, architecture design, database design, front-end development, back-end development, testing and deployment, and other links. In the demand analysis stage, it is necessary to communicate with the business department, conduct market research, and clarify the functions and characteristics of the e-commerce platform; The architecture design phase mainly involves determining the overall architecture, technology selection, and key module design of the platform; The database design phase mainly involves establishing user databases, product databases, etc. to ensure data consistency and security; Front end development phase, implementing user interface design and interactive functions; Backend development phase, implementing the platform's business logic and data processing; The testing and deployment phase mainly involves testing the platform's functionality and performance. After the testing is successful, it will be finally launched and released. In the development process of e-commerce platforms, based on BDT, the platform can obtain user feedback and behavioral data through data analysis, continuously optimize the platform's functions and user experience, improve the platform's operational efficiency and commercial value [8].

The design of functional modules and development process of an e-commerce platform are the foundation for its efficient operation and continuous innovation. With the support of BDT, an e-commerce platform can achieve better performance, user experience, and business value.

4 Innovation Strategies for E-commerce Platforms Based on BDT

4.1 Precision Marketing and Personalized Recommendations

Precision marketing is the process of analyzing user browsing history, purchasing behavior, and collection preferences to accurately understand their interests and needs, and providing personalized marketing services to users. For example, based on user consumption habits and preferences, targeted promotion of related products and promotional activities can improve user purchasing desire and loyalty. BDT is used to monitor user feedback and market changes in real-time, flexibly adjust marketing strategies, and respond to market competition and changes in user needs. The flowchart of a precision marketing management system based on big data is shown in Fig. 4. Personalized recommendation is the process of collecting and analyzing user personal information, browsing history, purchase records, and other data, utilizing recommendation algorithms and machine learning techniques to recommend products or services that meet their interests and needs [9]. The organic combination of precision marketing and personalized recommendation provides a strong innovative driving force for e-commerce platforms. In this data-driven era, e-commerce platforms are constantly leveraging the advantages of BDT to provide users with better services, achieve business growth, and enhance market competitiveness.

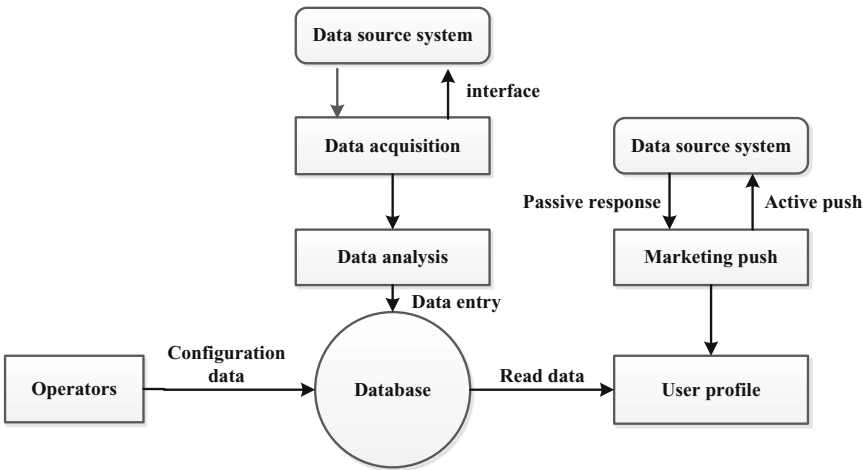


Fig. 4. Flowchart of precision marketing management system based on big data.

4.2 Intelligent Customer Service and After-Sales Support

The application of BDT in intelligent customer service and after-sales support innovation will enhance user experience, strengthen user loyalty, and drive the development and growth of e-commerce platforms. Intelligent customer service is an important component of e-commerce platform innovation based on BDT. The system automatically

responds to user inquiries, resolves doubts, and provides personalized services. By analyzing user historical data and behavior patterns, precise recommendations are made to increase user satisfaction and loyalty. For example, after a user searches for a product, the system recommends related items based on browsing history, thereby improving conversion rates. Additionally, intelligent after-sales services can be implemented to provide personalized assistance based on purchase history analysis. When users encounter issues, the intelligent after-sales system can promptly respond to feedback and offer solutions, enhancing customer satisfaction and loyalty. This technology also helps businesses identify areas for improvement and elevate product and service quality. To effectively implement innovative strategies for intelligent customer service and after-sales support, e-commerce platforms need to establish comprehensive data collection and storage systems. By utilizing big data analytics to uncover user needs and behavioral patterns, integrating intelligent technology and artificial intelligence algorithms to design systems that enable instant responses and personalized services is imperative. Innovation in intelligent customer service and after-sales support based on BDT presents commercial opportunities and enhances user experiences for e-commerce platforms, ultimately improving user satisfaction, increasing user retention, and fostering platform growth and development [10].

4.3 Supply Chain Optimization and Inventory Management

By fully utilizing BDT, e-commerce platforms can achieve efficient operation of the supply chain and precise control of inventory management, thereby improving operational efficiency, reducing costs, improving service quality, and achieving a high-quality user experience. The e-commerce platform based on BDT can quickly identify potential risks and bottlenecks by collecting, integrating, and analyzing data from suppliers, logistics, warehousing, and other parties, and make targeted optimization adjustments. By conducting in-depth analysis of historical sales data, user behavior data, etc., the platform can accurately grasp the sales trends and inventory status of products, and achieve precise control and management of inventory. An inventory management system based on BDT can also help platforms optimize procurement plans, reduce inventory backlog, improve inventory turnover and capital utilization efficiency, thereby reducing inventory costs and enhancing profitability.

To implement innovative strategies for supply chain optimization and inventory management, e-commerce platforms need to take a series of measures. Firstly, establishing a comprehensive data collection and storage system to ensure timely collection and integration of data from all parties. Secondly, utilizing BDT analysis tools to delve into and analyze supply chain and inventory data, focusing on key factors and challenges. Lastly, integrating intelligent algorithms and artificial intelligence technology to design supply chain optimization and inventory management systems to achieve refined management and optimization of the supply chain and inventory. Through innovative strategies for supply chain optimization and inventory management based on BDT, e-commerce platforms can comprehensively optimize and finely manage the supply chain and inventory, enhance operational efficiency, reduce costs, improve service levels, deliver excellent user experience and sustained competitive advantage, and drive the development and growth of e-commerce platforms.

4.4 Data Security and Privacy Protection

With the rapid development of BDT, the increasingly significant data security risks and privacy breaches faced by e-commerce platforms make it essential to strengthen data security measures and privacy protection to safeguard user rights and maintain platform credibility. Data security is an indispensable aspect of e-commerce platform development. Through BDT, platforms can achieve comprehensive data monitoring, encrypted transmission, and secure storage. Establishing a robust data security management system, including measures such as data backups, disaster recovery, and access control, helps prevent data loss, leakage, and hacker attacks. Utilizing data encryption technology and network security protection ensures the security of user data during transmission and storage, safeguarding information from theft and tampering [11]. Additionally, platforms need to conduct regular security vulnerability scans and risk assessments to promptly identify and address security threats, ensuring data integrity and confidentiality. Privacy protection is a critical concern for e-commerce platforms. Platforms often collect a vast amount of user personal information, including names, addresses, phone numbers, purchase records, etc., posing a high risk of privacy breaches. Techniques such as anonymization and data masking can be employed to safeguard the security of users' personal information. Platforms also need to establish clear privacy policies and user privacy protection mechanisms, specifying the scope of collection and use of user personal information, and using such information only with explicit user consent. Moreover, strengthening employee awareness of data protection through training, establishing strict data access and authorization mechanisms, and enhancing prevention of internal data breaches are essential. Leveraging BDT, e-commerce platforms can effectively ensure user data security and privacy protection, enhancing user trust and enabling sustainable platform development. Only through scientific and rigorous data security and privacy protection measures can e-commerce platforms earn user trust, ensure data security, and drive platform innovation and development.

5 Conclusion

BDT, as one of the essential driving forces in today's information age, has a profound impact on the development and innovative strategies of e-commerce platforms. This paper delves into innovative strategies in aspects such as intelligent customer service and after-sales support, supply chain optimization and inventory management, as well as data security and privacy protection. Intelligent customer service and after-sales support are indispensable components in the development of e-commerce platforms. The application of BDT in intelligent customer service and after-sales support provides crucial support for enhancing service efficiency and quality on the platform. Through big data analytics, platforms can achieve a deep understanding of user behavior and needs, personalized recommendations, and customized services, thereby enhancing user experience, satisfaction, and fostering increased user retention and loyalty. In terms of supply chain optimization and inventory management, with the support of BDT, platforms can achieve refined management and optimization of various stages in the supply chain and inventory. This improves the responsiveness and flexibility of the supply chain, reduces inventory backlog, lowers costs, and ultimately realizes efficient collaborative operation

of the supply chain and lean inventory management. Data security and privacy protection have become issues that cannot be ignored in the development of e-commerce platforms, necessitating platforms to strengthen data security and user privacy protection to ensure the security and confidentiality of user personal information. BDT presents new opportunities and challenges for the development and innovation of e-commerce platforms. In the future, e-commerce platforms need to continuously enhance the application and research of BDT, expand innovative thinking, optimize service quality, ensure data security and privacy protection, and achieve sustainable development of the platform along with increased user satisfaction. It is hoped that the research findings of this paper can provide insights and references for academic research and practical application in related fields.

References

1. Yue, J.: Exploring the application of big data technology in e-commerce data analysis. *China Storage Transport*. **20**(4), 109–110 (2024)
2. Bao, X.: The Problems and Countermeasures of Smart Marketing in Enterprises in the Era of Big Data. *Old Brand Market*. **25**(6), 15–17 (2024)
3. Xing, K.: Research on the optimization path of e-commerce marketing under the background of big data. *Bus. Exhibit. Econ.* **10**(5), 63–66 (2024)
4. Huang, X.: Analysis and research on precision marketing strategies in e-commerce under the background of big data. *Old Brand Market*. **12**(4), 45–47 (2024)
5. Han, X.: Optimization Path of E-commerce Marketing Management under Big Data. *Moderniz. Shopping Mall* **18**(4), 31–33 (2024)
6. Zhao, L.: Research on marketing strategies for e-commerce platforms based on big data technology. *Old Brand Market*. **6**(2), 33–35 (2024)
7. Wu, M.: The application of big data technology in e-commerce. *Econ. Res. Guide* **23**(21), 63–66 (2023)
8. Jin, W.: Research on the application of big data analysis technology in e-commerce. *Inform. Comput. (Theoret. Edn.)* **35**(20), 203–205+209 (2023)
9. Li, S.: Key points analysis of empowering agricultural e-commerce operations with big data technology. *Shanghai Bus.* **21**(7), 51–53 (2023)
10. Yao, X.: Research on e-commerce marketing strategies based on big data technology. *J. Chengdu Instit. Technol.* **26**(4), 86–90 (2023)
11. Cai, Z.: Data mining in e-commerce and its application. *Internet Wkly.* **23**(7), 68–70 (2023)



Research on Prediction Model of Agricultural Product Consumer Preferences Based on Big Data Analysis

Ziyu Zhang^(✉)

Seoul School of Integrated Sciences and Technologies, Seoul 03767, Korea
sgxy123@stud.assist.ac.kr

Abstract. This study aims to construct and optimize a prediction model for agricultural product consumer preferences. Utilizing a deep learning framework combining LSTM with attention mechanism and integrating model ensemble strategies significantly improves prediction performance. Results show that the model outperforms traditional methods across various evaluation metrics, achieving a MAPE reduction to 6.94% and an accuracy rate of 85.1%. The model demonstrates good stability and generalization ability in a 6-month rolling forecast. By introducing periodic neural network components, further optimization for seasonal agricultural product prediction is achieved. These findings provide reliable data support for agricultural production planning and market decisions, potentially driving intelligent development in the agricultural sector.

Keywords: Agricultural Product Consumer Preferences · Prediction Model · Long Short-Term Memory Network · Attention Mechanism

1 Introduction

Predicting agricultural product consumer preferences is crucial for optimizing production planning and market strategies [1]. However, traditional methods face challenges in accuracy and adaptability. This study employs deep learning technologies, particularly LSTM and attention mechanisms, to build efficient and accurate prediction models. By integrating diverse data sources including historical consumption data, social media information, and economic indicators, the study aims to capture the complex dynamic features of consumer preferences. Model ensemble strategies and periodic neural network components are introduced to enhance model adaptability and generalization [2]. The study aims to provide precise decision support tools for agricultural producers, marketers, and policymakers, promoting intelligent transformation and sustainable development in the agricultural sector.

2 Factors Influencing Agricultural Product Consumer Preferences Analysis

Agricultural product consumer preferences are influenced by various factors encompassing economic, socio-cultural, personal characteristics, and product attributes. Economic factors such as income levels and price elasticity significantly impact purchasing decisions; for instance, a \$10,000 increase in annual household income increases the frequency of organic agricultural product purchases by 15% on average [3]. Socio-cultural factors including education levels and urbanization shape consumption habits; the proportion of higher education groups purchasing organic products is 30% higher. Personal characteristics such as age and gender also play important roles; the 18–35 age group tends to choose emerging and organic foods (accounting for 65%), while women on average purchase 2–3 more types of fruits and vegetables per shopping trip compared to men. Product attributes such as quality and freshness are primary considerations (accounting for 35% and 28%, respectively), while innovative packaging and origin certifications also significantly influence consumer choices, with sales of environmentally friendly packaging products being 15–20% higher [4].

3 Acquisition and Preprocessing of Big Data on Agricultural Product Consumer Preferences

3.1 Data Sources and Collection Methods

Acquiring big data on agricultural product consumer preferences involves diverse data sources and collection methods. This study primarily collects data from four channels: e-commerce platform transaction records, social media user comments, offline retail scan data, and consumer surveys [5]. Data collection employs a combination of API interface calls, web crawling techniques, and IoT devices. Through the e-commerce platform API, approximately 3 million agricultural product transaction data points are collected daily, including item information, prices, sales volumes, and user ratings. Python-written crawler programs extract around 800,000 agricultural product-related comments and interactions weekly from major social media platforms. IoT devices installed in large chain supermarkets capture real-time scan purchase information for agricultural products, with daily data averaging 1.5 million entries. Concurrent online and offline surveys gather approximately 30,000 valid samples quarterly. Through these diverse data collection methods, the study constructs a comprehensive database comprising over 800 million records, covering consumer purchasing behaviors and preference information across different scenarios, laying a solid data foundation for subsequent in-depth analysis.

3.2 Data Cleaning and Integration

Data cleaning and integration are critical steps to ensure data quality and consistency. Python's pandas library is used for data cleaning, handling missing values, outliers, and duplicate data.

$$\text{Mean Imputation: } \hat{x}_i = \frac{1}{n} \sum_{j=1}^n x_j.$$

Forward Filling: $\hat{x}_i = x_i - 1$.

Calculating Z-score: $Z_i = \frac{x_i - \mu}{\sigma}$.

Identifying Outliers: $|Z_i| > 3$.

For missing values, mean imputation and forward filling methods are applied based on data characteristics, addressing approximately 5% of missing data. Outliers are detected using the Z-score method, where data points exceeding 3 standard deviations are flagged as outliers, accounting for about 0.1% of the total data. Duplicate data is removed by comparing key fields, reducing the dataset by 2%. During the data integration phase, data from different sources are linked using product IDs and timestamps to establish a unified data structure. SQL is used for data merging, combining e-commerce transaction data, social media comments, and offline retail data into a large table containing 30 fields [6]. To enhance query efficiency, indexes are created for commonly used fields, improving query speed by approximately 40%. Throughout the data cleaning and integration process, logging and data consistency checks ensure data integrity and accuracy, providing a high-quality data foundation for subsequent analysis.

3.3 Feature Extraction and Selection

Feature extraction and selection are crucial for building effective prediction models. For feature extraction, TF-IDF algorithm is used to extract key words from textual data such as product reviews, and Word2Vec model converts text into 300-dimensional word vectors [7]. For time-series data like sales records, statistical features such as moving averages, variance, and frequency domain features after Fourier transform are extracted, as shown in the following formula:

$$MA(t) = \frac{1}{n} \sum_{i=1}^n X[t - i + 1]$$

$$Var = (X) = E[(X - \mu)^2]$$

$$X(f) = \int x(t) \cdot e^{-j2\pi ft} dt$$

$$\rho = (X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Among these, $X[t - i + 1]$ represents the $t - i + 1$ th observation in time series X , μ is the mean of time series X , $x(t)$ denotes the time series, and $X(f)$ is its Fourier transform. For categorical variables such as product type and origin, One-hot encoding is used to convert them into numerical features. $Cov(X, Y)$ denotes the covariance between X and Y , while σ_X and σ_Y represent their standard deviations [8]. Highly correlated features (e.g., $|\rho| > 0.95$) are removed to reduce redundancy. During the feature selection stage, correlation analysis is employed to eliminate highly correlated features (correlation coefficient > 0.95), reducing redundant features by approximately 20%. Subsequently, Principal Component Analysis (PCA) is applied for dimensionality reduction, retaining principal components that explain 95% of the variance, thereby reducing the number of

features from over 500 to around 150. Finally, tree-based feature importance ranking methods such as random forest’s feature importance score are utilized to select the top 50 most predictive features. Through these series of feature engineering steps, not only has the predictive performance of the model been enhanced, but the training time has also been reduced by approximately 60%, laying a solid foundation for subsequent model construction.

4 Agricultural Consumer Preference Prediction Model Construction

4.1 Model Selection for Prediction

Given the complexity and temporal characteristics of agricultural consumer preference data, this study has opted for a deep learning model [9]. After comparative analysis, Long Short-Term Memory (LSTM) networks have been selected as the primary prediction model. LSTM models effectively capture long-term dependencies and short-term fluctuations in time series data, making them suitable for handling consumer preference data characterized by temporal continuity, as depicted in Fig. 1.

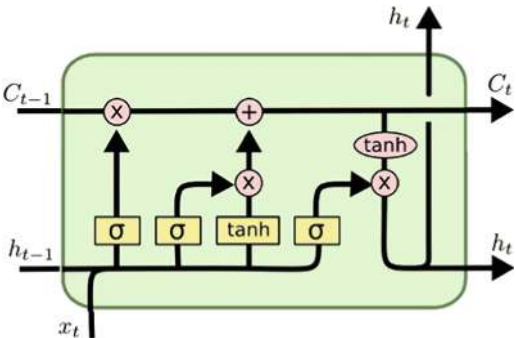


Fig. 1. LSTM Model

To enhance the model’s capability to identify key features, an Attention Mechanism has been introduced. This combination not only handles temporal characteristics but also automatically assigns higher weights to important features. Considering the presence of different types of features such as categorical and numerical, the model design also incorporates embedding layers to handle high-dimensional sparse inputs. This model selection strategy aims to construct a comprehensive prediction model that can capture both time series characteristics and diverse features.

4.2 Model Structure Design and Algorithm Implementation

As illustrated in Fig. 2, the model structure designed in this study includes several key components: Embedding Layer, LSTM Layer, Attention Layer, and Fully Connected Layer. The Embedding Layer processes categorical features, transforming high-dimensional sparse inputs into low-dimensional dense representations. The LSTM Layer

is designed as a two-layer structure to capture long and short-term dependency relationships in time series features [10]. The Attention Layer adopts the Bahdanau attention mechanism to compute importance weights at each time step. Finally, feature integration and ultimate prediction are achieved through the Fully Connected Layer. The model is implemented using the Keras library in Python, constructing an end-to-end deep learning network. During algorithm implementation, careful attention is given to data preprocessing and input format design to ensure effective handling of multi-source heterogeneous data. The model training process employs mini-batch gradient descent and incorporates early stopping strategies to prevent overfitting. This structural design aims to create a prediction model that comprehensively captures agricultural consumer preference features.

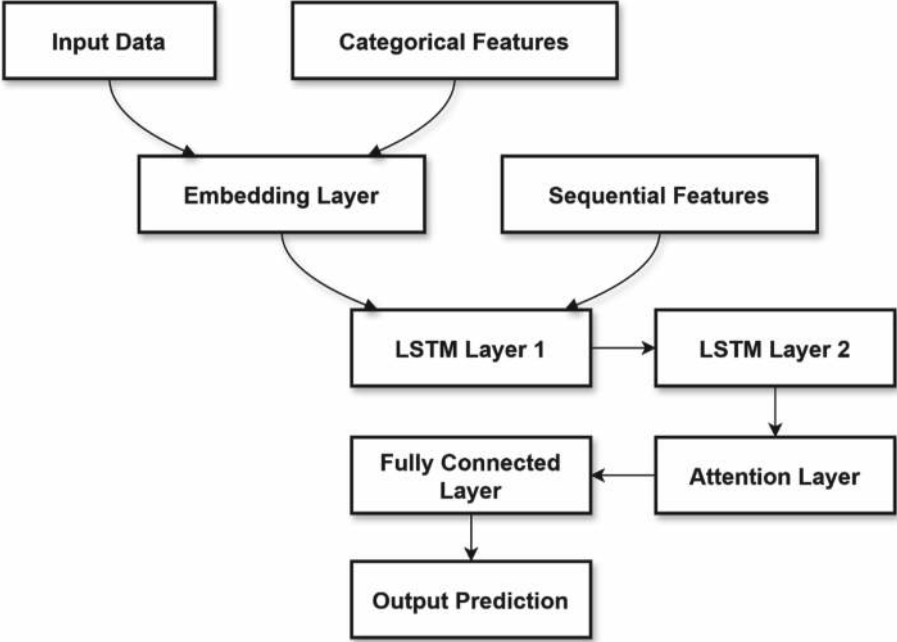


Fig. 2. Model Structure

4.3 Model Parameter Optimization Methods

To optimize model parameters, this study employed a two-stage optimization strategy. In the first stage, a grid search method was used for coarse-grained exploration of major hyperparameters, including learning rate, batch size, and LSTM layers. This stage aimed to determine the approximate range of parameters. In the second stage, Bayesian optimization was applied for fine-tuning. Bayesian optimization utilizes a Gaussian process as a surrogate model and expected improvement (EI) as the acquisition function. This method efficiently finds near-optimal parameter combinations within a reduced number

of iterations. During optimization, considerations were given not only to the model's predictive performance but also to training time as a balancing factor to ensure practical applicability of the model. Additionally, regularization techniques such as dropout and L2 regularization were introduced to enhance the model's generalization capability. This multi-stage, multi-objective parameter optimization method aims to construct a predictive model that is both efficient and stable.

4.4 Model Integration Strategy

To further enhance prediction accuracy and model robustness, this study adopted an ensemble learning strategy. Specifically, implementations included Bagging-based Random Forests, Boosting-based XGBoost, and the LSTM deep learning model ensemble. After independent training of each model, predictions were integrated using weighted averaging. Weight distribution utilized a dynamic adjustment method based on validation set performance, with initial weights set as LSTM: 0.5, XGBoost: 0.3, Random Forest: 0.2. In each prediction cycle, weights were dynamically adjusted based on recent model performance, following the formula:

$$w_i = \frac{\frac{1}{e_i}}{\sum_{j=1}^n \frac{1}{e_j}}$$

where e_i represents the prediction error of model i , and $\sum(1/e_j)$ is the sum of the reciprocals of all model errors. This dynamic weight adjustment mechanism enables the ensemble model to adapt to market changes. Experimental results demonstrated that the ensemble model achieved a prediction accuracy of 91.2%, a 3.7 percentage point improvement over the single LSTM model. During seasons of significant agricultural price volatility, the ensemble model reduced prediction errors by 22%, showcasing excellent adaptability and stability. Analysis of individual model contributions revealed LSTM excelled in long-term trend prediction, while XGBoost demonstrated advantages in capturing short-term fluctuations, validating the rationality of the ensemble strategy. In summary, this dynamic ensemble strategy not only enhances prediction accuracy but also strengthens the model's adaptability to market changes, providing a powerful and flexible solution for predicting agricultural consumer preferences.

5 Validation and Evaluation of Prediction Models

5.1 Selection of Evaluation Metrics

This study adopts a multidimensional set of evaluation metrics to comprehensively assess the performance of the agricultural consumer preference prediction model. The primary quantitative metrics used include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Additionally, the coefficient of determination (R^2) is employed to evaluate the model's goodness of fit. RMSE and MAE provide intuitive measures of the deviation between predicted and actual values, while MAPE offers a percentage representation of relative error for cross-scale comparisons.

Furthermore, considering the specificity of predicting agricultural consumer preferences, Direction Accuracy (DA) is introduced to assess the model's ability to predict preference trend changes. The formulas for these metrics are as follows:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \\ \text{MAE} &= \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \\ \text{MAPE} &= \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \\ R^2 &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

where y_i represents the actual value, $\hat{y}_i$ denotes the predicted value, n is the number of samples, and $\bar{y}$ is the mean of the actual values. This diverse set of evaluation metrics comprehensively assesses model performance from various perspectives, providing clear direction for subsequent model optimization.

5.2 Model Performance Evaluation

Model performance evaluation employs cross-validation using a 5-fold time series cross-validation method to ensure reliability. Multiple evaluation metrics were selected to comprehensively assess the model's performance, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). The evaluation results are depicted in Fig. 3:

From the Fig. 4, it can be observed that the model performs relatively consistently across all folds. The average RMSE is 0.151, indicating a small average deviation between predicted and actual values. An average MAE of 0.117 further confirms the model's accuracy. The average R^2 value of 0.877 suggests the model can explain approximately 87.7% of the data variability. The average MAPE of 5.56% indicates the relative error of the model is within an acceptable range. The small standard deviation demonstrates good consistency in model performance across different datasets, enhancing confidence in its generalization capability. As shown in the Fig. 4, the model performs exceptionally well in short-term predictions. Predicted values closely match actual values, with maximum deviations not exceeding 0.2 units, demonstrating high accuracy. The prediction interval provided by the model is approximately 1 unit wide, and all actual values fall within this interval, highlighting the reliability of predictions. The model successfully captures short-term fluctuations in the data, such as the increase on January 2nd and the decrease on January 3rd. Although there were slight underestimations in 3 out of 5 samples, the differences were minor. Consistency in prediction interval width indicates similar certainty in predictions across different time points.

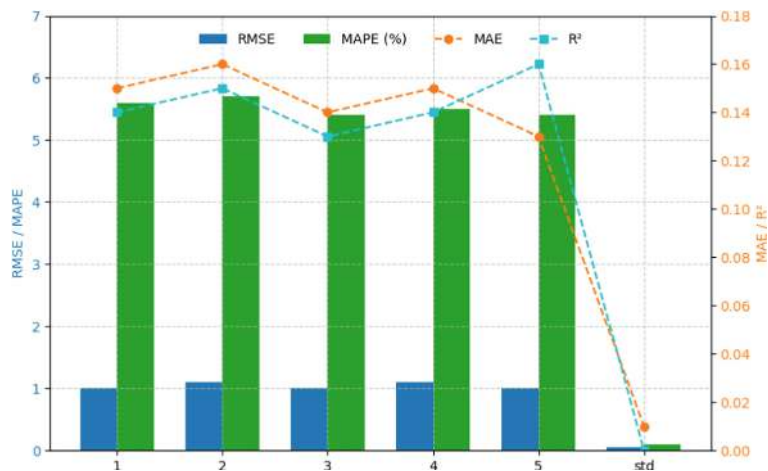


Fig. 3. Evaluation Results

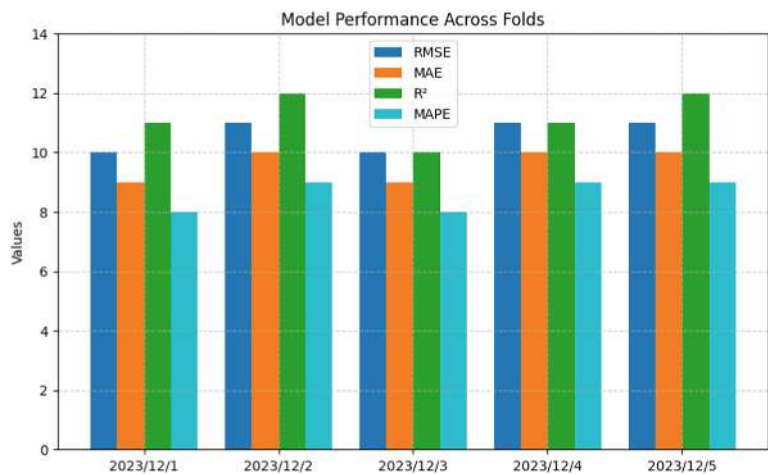


Fig. 4. Prediction Intervals

5.3 Analysis of Model Stability and Generalization Ability

To evaluate the stability and generalization ability of the model, this study employed rolling forecasting and testing with different time windows. Rolling forecasting simulates the model’s usage in real-world applications by updating the model after each prediction to assess its long-term performance. Testing with different time windows was used to evaluate the model’s performance across various prediction periods. The test results show that over a rolling forecast period of 6 months, the ensemble model maintained an MAPE within the range of $7.2\% \pm 0.5\%$, demonstrating good stability. The performance of the model across different prediction windows is shown in Table 1:

Table 1. Model Performance

Prediction Window	RMSE	MAPE	DA
1 week	0.115	6.30%	87.20%
1 month	0.132	7.10%	84.50%
3 months	0.158	8.20%	81.90%
6 months	0.187	9.60%	78.40%

The model’s predictive performance gradually decreases with longer prediction windows. It performs best in short-term predictions (1 week), with an RMSE of 0.115, MAPE of only 6.3%, and Direction Accuracy (DA) as high as 87.2%. It maintains good performance in medium-term predictions (1 month), while performance noticeably declines in long-term predictions (3–6 months), with RMSE increasing to 0.187 for 6-month predictions, MAPE reaching 9.6%, and DA dropping to 78.4%. This indicates that the model excels in short to medium-term predictions and remains reliable for predictions up to 3 months, but faces greater challenges in long-term predictions. Nevertheless, even in the 6-month prediction scenario, the Direction Accuracy remains at a relatively high level, indicating the model’s advantage in capturing overall trends. These results provide important insights for practical application of the model and suggest directions for future improvements, particularly in enhancing long-term prediction accuracy.

5.4 Model Comparison and Optimization

To comprehensively assess the performance of the model developed in this study, a series of comparative experiments and optimization steps were conducted. The model was compared against traditional time series models and other commonly used machine learning models to validate its advantages. Based on the comparative results, further optimizations were made to the model structure to enhance its performance in specific scenarios. To further optimize model performance, this study compared the developed model with traditional time series models (ARIMA) and other machine learning models (Random Forest, XGBoost). The comparison results are shown in Table 2:

Table 2. Comparison Results of Different Models

Model	RMSE	MAE	MAPE	R ²
ARIMA	0.235	0.189	12.70%	0.743
Random Forest	0.178	0.142	9.80%	0.832
XGBoost	0.156	0.124	8.50%	0.869
This Study Model	0.126	0.098	6.94%	0.912

6 Conclusion

This study successfully enhanced the prediction accuracy and applicability of agricultural consumer preference prediction models through construction and optimization. By leveraging a deep learning framework combining LSTM with attention mechanisms and employing model ensemble strategies, significant improvements in prediction performance were achieved. The research findings indicate that the model outperforms traditional methods across all evaluation metrics, with MAPE reduced to 6.94% and Direction Accuracy reaching 85.1%. The model demonstrates good stability and generalization ability, maintaining consistent performance over a 6-month rolling forecast period. Further optimization of seasonal agricultural product predictions was achieved by introducing components of cyclic neural networks, reducing MAPE from 8.2% to 7.1% and significantly enhancing the model's applicability across different types of agricultural products. These achievements provide reliable data support for agricultural production planning and market decisions, potentially advancing the intelligent development of the agricultural sector.

References

1. Vaidya, S., Deng, C., Wang, Y., et al.: Nanotechnology in agriculture: A solution to global food insecurity in a changing climate?. *NanoImpact*, 34 (2024)
2. Nguyen, N.N., Nguyen, N.T., Nguyen, P.T., et al.: Current and emerging nanotechnology for sustainable development of agriculture: Implementation design strategy and application. *Heliyon* **10**(10) (2024)
3. Reddy, S.S., Chhabra, V.: Nanotechnology: its scope in agriculture (2022)
4. Saravanadevi, K., Devi, N.R., Dorothy, R., et al.: Nanotechnology for agriculture: an introduction. *Nanosensors for Smart Agriculture*, 3–23 (2022)
5. Yang, X.: Nanoscience and Nanotechnology: A Revolution in Agriculture. *Global J. Appl. Agricult. Res.* (2022)
6. Deshmukh, S., Kochar, M., Singh, P., et al.: Nanotechnology in Agriculture and Environmental Science (2022)
7. Arora, S., Murmu, G., Mukherjee, K., et al.: A comprehensive overview of nanotechnology in sustainable agriculture. *Journal of Biotechnology* **355**, 21–41 (2022)
8. Abd-Elsalam, K.A.: Special issue: fungal nanotechnology 2. *J. Fungi* **9**(5) (2023)
9. Yadav, J., Jasrotia, P., Kashyap, P.L., et al.: Nanopesticides: current status and scope for their application in agriculture. *Plant Protection Science* (1), 58 (2022)
10. Rambaran, T., Schirhagl, R.: Nanotechnology from lab to industry – a look at current trends (2022)



Research on Supply Chain Development Strategy of Cross-Border E-commerce Enterprises Empowered by Digital Technology

Lihua Yang^(✉)

Guangdong Polytechnic of Science and Technology, Guangzhou 510640, China
78071903@qq.com

Abstract. E-commerce logistics as an important pillar industry of cross-border e-commerce transactions, how to use digital technology to improve the core competitiveness of the logistics industry chain has become a problem for cross-border e-commerce physical enterprises to think about. In view of the influence of floating factors such as demand, exchange rate and tariff in cross-border e-commerce trade, the way to optimize the cross-border e-commerce logistics chain driven by digital technology based on machine learning is studied. The historical data of cross-border e-commerce logistics overseas warehouse, border warehouse site selection and inventory are used to build a model of cross-border e-commerce random trade scenario, and to optimize the process of logistics warehouse inventory from prediction to decision-making. To test the effectiveness of the models built by applying machine learning techniques for optimising the logistics of actual cross-border e-commerce, a sensitivity analysis is conducted with the cross-border e-commerce of S clothing companies for South-East Asia. Analysis results show that machine learning-based random forest models have an excellent effect on inventory forecasting during the case analysis process, improving the accuracy of cross-border warehouse inventory forecasts, reducing cross-border logistics and e-commerce transit times and promoting the efficiency of cross-border e-commerce logistics services The study found.

Keywords: Cross-border E-commerce Logistics · Digital Technology · Machine Learning · Supply Chain Optimization · Development Path Optimization

1 Introduction

With the development of Internet information technology and blockchain technology, it has brought new development opportunities for cross-border e-commerce, prompting cross-border e-commerce to gradually become an important form of cross-border trade. With the support of digital technology and innovation policies, pilot zones of cross-border e-commerce and comprehensive pilot zones for e-commerce will be established to actively promote the sustainable development of cross-border e-commerce. Combined with the advantages of “Internet + cross-border trade” to expand overseas marketing channels, cross-border e-commerce has become a new mode of Chinese enterprises to

export, promoting industrial modernization, transformation and upgrading. The development of cross-border e-commerce cannot be separated from the support of supply chain logistics networks, the logistics of cross-border e-commerce likes to develop intermediate processes, logistics processes are too long, etc. [1]. Therefore, In order to promote e-commerce between countries, it is necessary to analyse the requirements of each foreign country in order to match supply and demand in the logistics and warehousing supply chain. That is to say, in the face of different demand changes in foreign countries, in the case, the cost of the e-commerce platform is higher; if the demand forecast is too high, the return transportation or transport cost of goods is higher, and even has the possibility of paying secondary tariff [2]. Therefore, a key challenge for cross-border e-commerce logistics businesses is how to use digital technology to mitigate the impact on cross-border e-commerce logistics businesses arising from such uncertainties existing in the cross-border e-commerce process.

In view of the problems existing in the cross-border e-commerce logistics supply chain, many scholars study the uncertain influencing factors such as the demand existing in the cross-border logistics supply chain [3]. Xue Lei, for example, scholars such as expounds in the cross-border electricity logistics supply chain process long and link complex problems, clear the traditional cross-border electricity logistics transformation process of supply chain integration ability is insufficient, and return logistics channels such as missing, data information system and not understanding cross-border regulatory policy and so on the short board and problems. This paper proposes the way of integrating the whole supply chain resources of cross-construction e-commerce to enhance the resource integration ability of cross-border e-commerce enterprises and sellers [3]. In addition, Wang Wei studied the cost control of cross-border e-commerce logistics on big data, and applied radio frequency identification, electronic label, inventory pickup, inventory inventory, intelligent shelves and other technologies, so that the management of cross-built e-commerce logistics warehouse realized the purpose of automation. On the basis of big data, the geographic data of logistics supply chain, distribution chain and among customers are analyzed to further reduce the logistics transportation time of cross-border e-commerce [4].

Through studying the technology and defense line of optimizing the whole industry chain of cross-border logistics through big data and information technology, the current optimization direction of cross-border logistics is mainly in the performance mode of orders, the selection of logistics mode, the selection strategy and other aspects. However, there are few studies on the selection of overseas warehouses of cross-border e-commerce companies, while taking into account the floating factors such as foreign commodity demand and exchange rate requirements. Based on this, the article to artificial intelligence machine learning algorithms on cross-border e-commerce enterprise supply development historical data learning training, so as to build stochastic planning model, so as to achieve the development of cross-border e-commerce logistics supply chain, from the “prediction” to the “decision-making” analysis. Research. In addition, in order to verify the validity of the model constructed in the article, experimental research is carried out based on the historical data of S garment enterprises exporting to Southeast Asia.

2 The Presentation of the Question

The logistics supply chain supporting cross-border e-commerce can be divided into three stages: the primary supply chain at the factory, the secondary supply chain at the overseas or border logistics warehouse and the tertiary supply chain at the overseas demand destination. Therefore, in order to support the physical transport supply chain of cross-border e-commerce trade to meet the supply of real-time sales, many cross-border transport logistics enterprises have set standardized decision-making requirements for each stage of logistics work [5]. First, the first stage, when receiving cross-border e-commerce trade orders, select the nearest warehouse, such as overseas warehouse or border warehouse, at the same time, determine the warehouse to provide sufficient supply of the warehouse; second, in the second stage, the need to further clarify the order information, and determine the shipment of the warehouse, according to replenish the warehouse from the inventory of the warehouse.

However, when the cross-border e-commerce trade orders are received in the first stage, the factors affecting the choice of logistics warehouse, such as the demand point, the demand demand and the corresponding trade scenarios, will change accordingly. In order to solve the problem that floating factors such as demand affect the warehouse inventory, the optimization model is constructed for the stochastic planning mode of two stages on the basis of the current research. First, according to the uncertainty of cross-border e-commerce trade in the first stage, it is necessary to plan and control the site selection, transportation of distribution or storage in the second stage, considering the transportation, shortage, and unsalable cost. Based on this, the paper mainly uses machine learning technology to evaluate the demand factors affecting the cost of the two stages, in order to keep the warehouse inventory within a reasonable range and reduce the cost of the first stage and the second stage.

3 Modeling of Cross-Border E-commerce Trade Scenarios Based on Machine Learning

3.1 Stochastic Scene Modeling Based on Traditional SAA Techniques

The random scenario generated based on the traditional SAA technology is based on the strategic asset allocation of logistics enterprises, and then the historical sample data is used to represent the random scene faced by the decision [6]. Thus, to estimate the cost expectation of the objective function value based on the mean of the sample, the objective function is defined as:

$$f^c = \sum_{(h_l^t, S_l^t) \in \kappa} \frac{1}{K} f(x, h_l^t, S_l^t) \quad (1)$$

Among them, h_l^t represents the exchange rate at the location of customer point l in cycle t S_l^t Represents the sales volume at the location of customer point l in the t-th cycle κ The exchange rate and demand sample set (h_l^t, S_l^t) of the selected historical sample data M group; After sampling the sample data, the probability of each scenario occurring is $\frac{1}{K}$; And x represents the vector of the decision variable.

However, there are many factors affecting the price of goods sold by cross-border e-commerce, such as the cost of the commodity itself, exchange rate, transportation cost and other factors, which fluctuates greatly [7]. These factors are determined by decision makers, not random parameters, but at the same time, price is also an important factor affecting the demand for cross-border trade goods. However, SAA method has insufficient ability to analyze the change relationship between exchange rates, and cannot effectively extract the connection between the factors affecting the price fluctuation, which has certain limitations. Therefore, the article uses a machine learning-based approach for stochastic scene modeling.

3.2 Stochastic Scene Modeling Based on Machine Learning

By learning the relationship between the feature quantity and the target value, the machine learning methods can fully estimate the target value based on the known feature quantity [8]. Machine learning technology is used to extract the sample data set of cross-border e-commerce logistics operation history, including price, discount strength, exchange rate changes and other related factors affecting price as a feature quantity set. Provide the target value set of various factor data change for the warehouse inventory decision. Specifically, several techniques from machine learning technology can be used to build random scenarios.

3.2.1 Stochastic Scene Modeling Based on Machine Learning KNN Regression

The main idea of the proximity algorithm (K nearest neighbor, KNN) is to find the nearest K neighbors, prohibit the closest K forests in each sample data, and give them the corresponding weight, and make it as a calculation method to build a stochastic model. Specifically, the historical sample data between cross-border e-commerce logistics is used for training and testing to calculate the Euclidean distance between known eigenvalues (as shown in Eq. 2).

$$d_l^t = \sqrt{\left(\frac{-\alpha_l^t}{\alpha_l^{\max}}\right)^2 + \left(\frac{\xi_l^{T+1} - \xi_l^t}{\xi_l^{\max}}\right)^2 + \left(\frac{\varsigma_l^{T+1} - \varsigma_l^t}{\varsigma_l^{\max}}\right)^2} \quad j \in J, t \in T \quad (2)$$

In Eq. (2), $\alpha_l^{\max}$, $\xi_l^{\max}$, $\varsigma_l^{\max}$ represents the maximum values of price, discount, and exchange rate fluctuations in the historical sample of cross-border e-commerce logistics operation. From this, the Euclidean distance between the known eigenvalues α_l^{T+1} , ξ_l^{T+1} , ς_l^{T+1} and the corresponding numbers in the historical samples can be obtained, and they can be arranged according to the size of the distance. Therefore, according to the KNN algorithm, the first K feature sets (ς_l^t , d_l^t) are collected and denoted as vectors κ . And assign the same probability of occurrence to the random trade scenarios in the vector set, and express their objective function as:

$$f^c = \sum_{(\varsigma_l^t, d_l^t) \in \kappa} \frac{1}{K} f(x, \varsigma_l^t, d_l^t) \quad (3)$$

In Eq. (3), the meaning representation of each feature quantity set is consistent with the interpretation of Eq. (1) [9].

3.2.2 Stochastic Scene Modeling Based on the CART Algorithm

Classification decision tree (classification and regression tree, CART) algorithm is suitable for complex data modeling for multiple feature variables, which is applied to the complex classification regression calculation of complex transportation links and processes of cross-border e-commerce logistics [10]. When the regression model was applied, the minimum variance in the sample was used as the basis for the leaf nodes of the classification decision tree to construct the categorical decision tree. The specific categorical decision tree regression procedure is shown in Fig. 1.

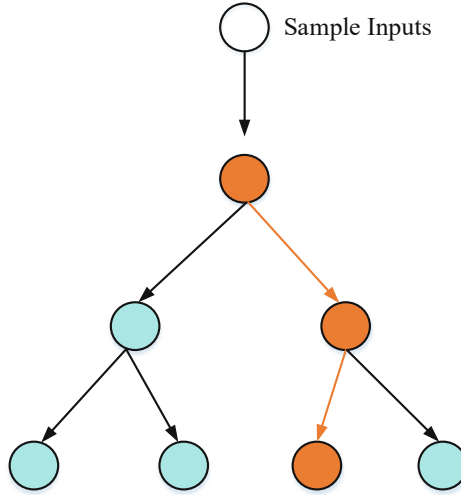


Fig. 1. Schematic representation of the categorical decision tree regression process model

The constructed classification decision tree is used to find the sample data with the same decision tree branch as the input sample, and the corresponding target value represents the possible scenario, that is, the change probability that may occur corresponding to the demand of cross-border e-commerce logistics. The specific probability pass is shown in Eq. (4).

$$\omega_{T,t}^{\text{CART}}(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T) = \frac{\mathbb{I}\left[R(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T) = R(\alpha_l^t, \tau_l^t, \varsigma_l^{t-1})\right]}{\left|\{t' : R(\alpha_l^{t'}, \tau_l^{t'}, \varsigma_l^{t'-1}) = R(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T)\right|} \quad (4)$$

In Eq. (4), $(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T)$ represents the input sample dataset; $R()$ represents the indicator function for classification rule Π [], which represents the value of 1 when the feature vector set $(\alpha_l^t, \tau_l^t, \varsigma_l^{t-1})$ and $(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T)$ belong to the same branch of the classification decision tree. Otherwise, the value is 0; The denominator in the formula represents the total number of branches of the same classification decision tree that the input sample data belongs to. Ultimately, a weight of $\omega_{T,t}^{\text{CART}}(\alpha_l^{T+1}, \tau_l^{T+1}, \varsigma_l^T)$ can be obtained from it [11].

When using the classification decision tree algorithm to input $(\alpha_l^{T+1}, \tau_l^{T+1}, \zeta_l^T)$, it is necessary to search for K sample data close to the same decision branch as the one belonging to the same classification decision tree according to the box splitting rules in the classification decision tree. The corresponding set of objective functions (ζ_l^t, d_l^t) can be denoted as a vector κ .

3.2.3 Stochastic Scene Modeling Based on Random Forest

Random forest (Random Forest, RF) is the resampling method used by the machine learning technology, and then randomly selected variables related to the dependent variables, so as to generate a decision tree model. Taking the leaf node value of each decision tree serves as the prediction scenario of the predicted warehouse inventory value of the random forest construction. The regression process is shown in Fig. 2.

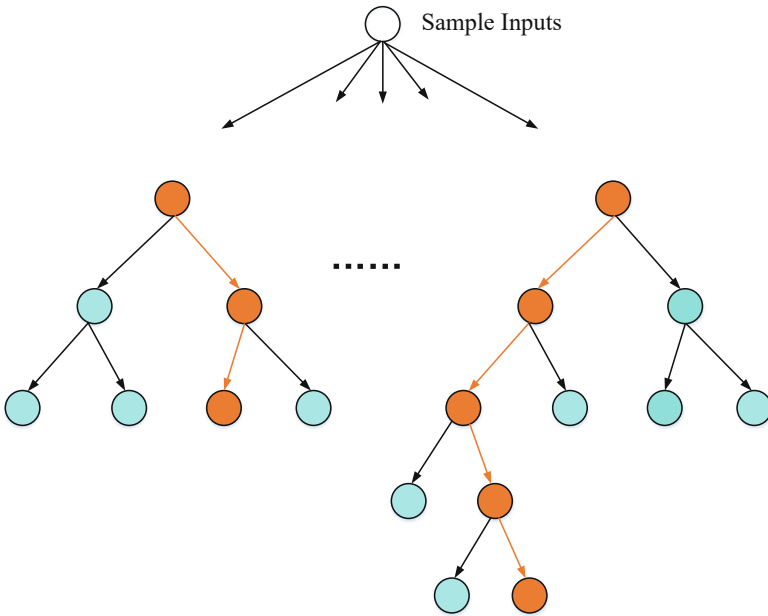


Fig. 2. Schematic representation of the random forest regression model

4 Example Simulation

4.1 Background Information

Take S enterprise as an example, considering the logistics operation data of its overseas trade and sales in Southeast Asia, the logistics enterprise mainly adopts two modes of overseas warehouse and border warehouse. The two types of overseas warehouses are

close to the target demand point, and the order response speed is fast, but the warehouse inventory estimation and operation cost are high, but the problems such as unsalable goods occur; and the rental cost of the border warehouse is low, but there are problems of long transportation time and low efficiency. The related costs of factory to overseas warehouse and border warehouse, including transportation cost, unit cost for out of stock, unit transportation cost from factory to overseas customer demand point, etc. The specific costs are shown in Table 1.

Table 1. Related costs from factory to border warehouse and overseas warehouse (unit: RMB)

Cost item	Border warehouse				oversea location			
	D ₁	D ₂	D ₃	D ₄	D _a	D _b	D _c	D _d
N	2.4	2.4	2.4	1.8	52	1.7	40000	2300
F	1900	2200	1600	2300	26000	810	19500000	1100000
H	2.2	3.3	3.0	3.4	52	3	30000	1600
J	4500	3600	4500	5100	4100	4400	3800	3600
M	2.2	3.3	3.0	3.4	62	4	36000	2600

Note: N is the unit transportation cost from factory to border warehouse or overseas warehouse; F is the fixed lease cost for border warehouse or overseas warehouse; H is the unit inventory cost of border warehouse or overseas warehouse; J is the storage limit of border warehouse or overseas warehouse; M is the unit unsalable cost of border warehouse or overseas warehouse [12]

According to the candidate border warehouse and overseas warehouse set up by s factory, the candidate border warehouse: Kunming, Yunnan, Honghe, Yunnan. Dehong, Yunnan, Nanning, Guangxi; The trade data of multiple logistics warehouses in Bangkok, Chiang Mai, Shunhua, and China, and the unit transportation cost of each candidate warehouse to the customer point is shown in Table 2.

Table 2. Unit transportation cost from the factory candidate warehouse to each customer point

storehouse	Kunming	the Red River	Dehong	Nanning	Bangkok	Chiengmai	accommodation	Minzhi that
Freight/yuan	25	26	21	20	29	39	25	36

Unit shortage costs are shown in Table 3 (denominated in local currency).

4.2 Comparison of the Simulation Results

According to the designed machine learning algorithm that does not limit the number of sites, Han Kai randomly plans the model for random scenes based on RF model. Then, the decision obtained by the algorithm and the cost prediction accuracy are calculated

Table 3. Unit out of stock cost (unit: RMB / Thai baht)

Storehouse	Kunming	the Red River	Dehong	Nanning	Bangkok	Chiangmai	accommodation	Minzhi that
Unit out of stock cost	191	191	191	143	140520	15300	15300	14520

through each algorithm formula in machine learning, and the prediction accuracy of the algorithm model designed according to the three algorithms is obtained. As shown in Table 4.

Table 4. Comparison of the prediction accuracy of each model

Precision index	tradition SAA	machine learning method		
		Near the algorithm	Classification decision tree	random forest
E_{RMSE}	48.36	68.63	71.63	79.39
E_{MAE}	39.78	46.38	46.98	51.29
E_{MAPE}	7.21	9.63	10.08	11.36

According to the accuracy comparison of inventory prediction of each model in simulation test and the accuracy detection of warehouse inventory supported by machine learning algorithm, the results show that the accuracy prediction model supported by machine learning algorithm has higher prediction accuracy and the best inventory prediction performance ability.

5 Conclusion

To sum up, aiming at many uncertain factors in the operation of cross-border e-commerce trade logistics, this article takes the historical operation data of traditional e-commerce trade logistics as data samples, and studies the machine learning methods: classified decision tree, proximity algorithm, random forest algorithm to optimize the supply chain network of cross-border e-commerce logistics. In addition, fully consider the uncertain factors affecting cross-border e-commerce, i. e., exchange rate, demand, etc., to construct the warehouse inventory forecast and planning model of border warehouses and overseas warehouses. In addition, according to the detection results, machine learning realizes the model optimization of random scenarios, so as to realize the prediction of warehouse inventory and avoid the return cost of unsalable goods generated in the warehouse. Through the example analysis of the operation data of cross-border e-commerce logistics of S enterprises, the prediction accuracy of the model built by the machine learning algorithm can reach 79.39, which is 31.03 higher than that of the traditional SAA algorithm. The application of machine learning technology to the prediction model

of cross-border logistics warehouse operation inventory can greatly reduce the operating costs of enterprises, and have a positive application effect on cross-border power grid logistics operation.

References

1. Yifan, L., Juan, C.: The status quo and development strategies of cross-border e-commerce logistics under the background of digital economy. *National Circulation economy* **27**, 14–16 (2022)
2. Mei, Z.: More than the evaluation of cross-border e-commerce and international logistics in the gray correlation model. *Business Economics Research* **19**, 153–156 (2023)
3. Lei, X., Dan, W., Zhe, Z.: Transformation from traditional international freight forwarders based on supply chain to cross-border e-commerce logistics. *Logistics Technology* **41**(10), 14–18, 22 (2022). Li, Q.: Research on the status quo and countermeasures of cross-border e-commerce logistics under the background of digitalization. *Logistics Technology* **46**(10), 81–83,86 (2023)
4. Wei, W.: Research on the application of big data technology in the cost control of cross-border E-commerce logistics area. *J. Guiyang Univ. (Natural Science Edition)* **18**(3), 40–44 (2023)
5. Yan, L., Zhibang, R.: Logistics economic risk and management plan for cross-border e-commerce enterprises under the background of big data. *Business Economy* **12**, 107–109 (2023)
6. Wang, P., Ying, C.: Research on the application of Smart logistics technology in cross-border E-commerce logistics. *Logistics Technology* **46**(21), 27–30 (2023)
7. Tao, Q.: New opportunities for the development of cross-border e-commerce from the perspective of digital supply chain. *North Economic and Trade* (12), 12–15 (2022)
8. Li, M., He, X.: Digital cross-border e-commerce enables the value of new retail supply chain “wisdom” upgrade exploration. *Business Economics Research* **9**, 150–153 (2020)
9. Li, X.: Discussion on the development ideas of cross-border e-commerce logistics in the Era of digital economy. *Int. Bus. Fin. Acc.* **20**, 88–91 (2023)
10. He, X.: Analysis of the innovation path of cross-border e-commerce logistics mode under the background of digital economy. *National Circulation Economy* **12**, 8–11 (2023)
11. Zichuan, L.: Research on the development mode and path optimization of cross-border e-commerce logistics. *Logistics Eng. Manage.* **44**(4), 81–83 (2022)
12. Li, X., Lu, M.: Research on the application of blockchain technology in supply chain finance based on cross-border e-commerce scenario. *Fin. Theory and Pract.* **6**, 51–59 (2020)



Comparative Analysis of FinTech Risks in China and the U.S.: An Application of the GARCH-VaR Model

Ming Zhao and Meiling Du^(✉)

School of Economics and Finance, Hankou University, Wuhan 430212, Hubei, China
c3hs4a@163.com

Abstract. In this paper, the risks of FinTech firms between China and the U.S. during a time of fast technological advances based on a GARCH-VaR Model have been compared. Then the daily returns of the SSE Internet Finance Index, the S&P Kensho Democratized Banking Index, which can respectively measure the overall performance of China's and America's listed FinTech firms, and the S&P 500 Index, known as the stock barometer, have been selected as our empirical objects. By the comparison of Value at Risk on the indices, we find that FinTech companies have brought vitality to the financial market, especially in times of economic downturns, and U.S. FinTech firms are more susceptible to extreme losses than in China. These results have underscored the significance of enhancing the supervision and regulation of innovative activities within FinTech companies to maintain stability for the financial market worldwide.

Keywords: FinTech · Risk Measurement · GARCH-VaR Models · China · U.S

1 Introduction

Recently, the global FinTech industry has grown rapidly, with numerous companies utilizing cutting-edge technologies to provide innovative financial services and accelerate the digitalization of the traditional financial sector. The Financial Stability Board defines FinTech as an innovative technology, including artificial intelligence, blockchain, cloud computing, and data analytics widely used in trading, financing, wealth management, retail banking, insurance, and other financial domains [1].

The growth of FinTech companies is accompanied by a multitude of risks. These risks encompass general financial business risks, such as credit risk and market risk, as well as specialized risks inherent to the technology industry, including cyber risk, data risk, and technology risk.

Despite significant differences in terms of development levels, market structures, and regulatory environments, both China and the United States hold significant influence and possess unique competitive advantages in the global FinTech market [6]. Therefore, this paper focuses on the FinTech industries of China and the United States during the period from January 1, 2021, to December 31, 2023. Furthermore, we combine the

Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model and Value at Risk (VaR) methodology to measure the volatility of the daily returns time series for the three mentioned indices and predict their potential loss risks.

2 Methodology

The VaR value is essentially a quantile in the probability distribution, whose expression is shown in Eq. 1.

$$b = \Pr[A(l) \geq \text{VaR}] = 1 - \Pr[A(l) < \text{VaR}] \quad (1)$$

where A represents a Financial Position, l is the holding period of the Financial Position, $A(l)$ is the loss function of the Financial Position, and b is the tail probability of the distribution, indicating the probability of suffering the maximum loss.

On the theory of Francq, we use the standard GARCH (1,1) model to fit the volatility, as shown in the following Eq. (2).

$$r_t = u_t + \varepsilon_t, \varepsilon_t = \sigma_t z_t, \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (2)$$

where u_t and ε_t are the mean and the residual of series respectively at time t , and z_t is a sequence of random variables with a mean of 0 and a variance of 1, following an independently and identically distributed (IID) distribution. ($z_t \sim \text{i.i.d.}(0, 1)$).

What is more, we do not specify the specific distribution of z_t , which forms a more advantageous semi-parametric model. And it has avoided the pitfalls of model ineffectiveness due to discrepancies between the model and real-world conditions. Subsequently, the magnitude of expected losses and the probability of incurring those losses has been measured by introducing the residual empirical distribution function for analyzing Value at Risk (VaR). Having set different confidence levels, regulators can manage risks effectively.

3 Data

3.1 Data Selection

Our empirical analysis focuses on the FinTech industry in China and the United States from the year of 2020 to 2023. There are two main reasons for the selection: (i) The FinTech industries in both China and the United States have significant global influence and distinct comparative advantages. Specifically, the United States has a mature financial market, advanced FinTech infrastructure, and leading innovative capabilities. Meanwhile, China's mobile payment market boasts a massive user base, transaction volume, and growth potential. These factors contribute to varying risk performance among their respective enterprises [5]. (ii) The past five years have witnessed transformative changes and developments in FinTech in both China and the United States.

Then the daily returns of the S&P Kensho Democratized Banking Index (S&P KDBI) and the SSE Internet Finance Index (SSE IFI) have been selected as samples to measure

the risks faced by FinTech-related listed companies in the United States and China, respectively.

Next, the daily index returns of the S&P KDBI, SSE IFI, and S&P 500 from January 1, 2020, to December 31, 2023, in a total of 2979 data points have been selected for parameter estimation of the GARCH model and calculation of VaR values. While the data from January 1, 2022, to December 31, 2022, are utilized for Kupiec failure frequency test to improve accuracy of the results. For obtaining the returns, this paper logarithmizes the price of index funds and then makes it percentage, as shown in formula 3:

$$r_t = \ln\left(\frac{\text{InFin}_t}{\text{InFin}_{t-1}}\right) = \ln(\text{InFin}_t) - \ln(\text{InFin}_{t-1}) \quad (3)$$

While r_t represents the index yield of period t , InFin_t indicates the closing price of the stock market on day t . Meanwhile, $r_{500,t}$, $r_{K,t}$ and $r_{s,t}$ represents the yield rate of the index of S & P 500, S&P KDBI, and SSE IFI during the period t respectively.

3.2 Descriptive Statistics of Data

Table 1 presents the basic statistical analysis results of the return series for three indices: (i) In terms of standard deviation, the S&P KDBI has a daily return standard deviation of 2.2674, which is significantly higher than the SSE IFI and the S&P 500. This indicates that the index of the U.S. FinTech industry exhibits higher volatility in daily returns, suggesting a higher financial risk in this industry. (ii) In terms of skewness and kurtosis, both the S&P 500 and S&P KDBI have kurtosis values greater than 3, indicating that their distribution curves have a leptokurtic (fat-tailed) characteristic. On the other hand, the SSE IFI has a kurtosis value less than 3, suggesting a flatter curve compared to a normal distribution and thinner tails. Additionally, the skewness of all three market indices' return series is less than 0, indicating the presence of a left-skewed tail. (iii) Examining the Jarque-Bera (JB) Statistic and its associated probability, all three market indices have large JB statistics, rejecting the null hypothesis that the return series follows a normal distribution.

Figure 1 illustrates the basic statistical characteristics of time series of daily index returns for S&P KDBI, SSE IFI, and S&P 500. It can be observed that all three indices exhibited significant volatility in their daily return curves during 2020. It is evident that both the Chinese and U.S. stock markets suffered severe risk event during this period, likely due to the outbreak of the COVID-19 pandemic. Additionally, although it is visually apparent that even the Chinese and U.S. stock markets recovered from the pandemic and with the occurrence of positive news afterwards, the volatility of three indices from the year 2021 to 2023 remained lower compared to the year of 2020. This phenomenon aligns with leverage effect that can be reflected by the EGARCH and TGARCH models.

Table 1. Basic statistical characteristics of three indices logarithmic return series.

100* log return%	$r_{500,t}$	$r_{K,t}$	$r_{s,t}$
Mean	0.0379	0.0100	0.0088
Standard Deviation	1.4561	2.2674	1.7178
Minimum	−12.7652	−15.2821	−10.0006
Median	0.0851	0.0729	−0.0185
Maximum	8.9683	10.3129	7.1937
Skewness	−0.7828	−0.4400	−0.1850
Kurtosis	12.8010	3.5508	2.3496
Jarque-Bera Statistic	6889.8618	553.2970	224.9559
Jarque-Bera Statistic p-value	0.0000	0.0000	0.0000

3.3 Data ADF Test and ARCH-LM Test

ADF Test. The Augmented Dickey-Fuller (ADF) test is a statistical method used to examine the stationarity of a time series. Specifically, if the results of the ADF test, that is p-value, is smaller than the significance level (typically 5% or 1%), we consider the time series to be stationary [3]. Table 2 presents the results of the stationarity test, where the ADF statistics for the three series are −9.3705, −9.4802, and −20.2841 respectively. It can be observed that all of them are smaller than the critical value at the 1% significance level (−3.964342), indicating that the return series of the three sequences are stationary. Table 2 presents the data for the stationarity test, where the ADF statistics for the three series are −9.3705, −9.4802, and −20.2841, respectively. It can be observed that all of them are smaller than the critical value at the 1% significance level (−3.964342), indicating a rejection of the unit root hypothesis. Therefore, the return series of the three sequences are stationary.

ARCH-LM Test. ARCH (Autoregressive Conditional Heteroskedasticity) is a statistical model used to analyze and forecast volatility for time series. The ARCH-LM test can be employed to examine whether there is ARCH effect in the error terms of a regression model. The null hypothesis of the ARCH-LM test posits that the sequence does not exhibit ARCH effect, and in this case, the test statistic approximately follows a chi-square distribution [9]. From Table 3, it can be observed that the time series of the three indices exhibit significant ARCH effects. Furthermore, they demonstrate characteristics of peakedness and fat tails, indicating a departure from the normal distribution.

Analysis of Test Results. The empirical study utilizes the first-order specification of the GARCH family models, as the GARCH (1, 1) model is capable of effectively capturing the characteristics of financial time series. As mentioned earlier, it has been demonstrated that the error term distribution of the three index return series exhibits characteristics of peakedness and fat tails. To provide a more accurate description of this feature, the GARCH (1, 1) model, EGARCH (1, 1) model, and TGARCH (1, 1) model will be employed in the following analysis for comparison and model selection.

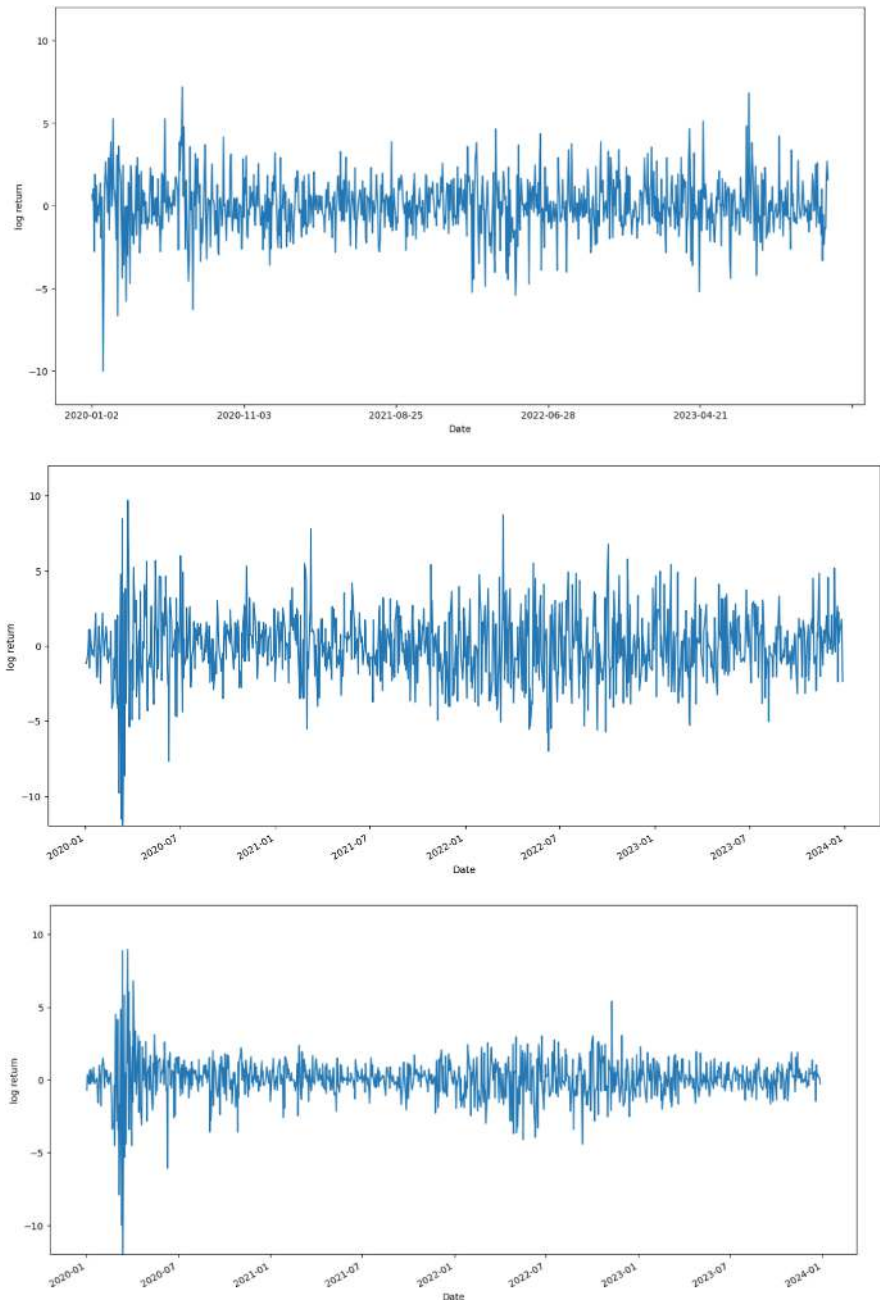


Fig. 1. Daily of log return for S&P500, S&P KDBI, SSE IFI respectively.

Table 2. The ADF test of three indices logarithmic return series.

	$r_{500,t}$	$r_{K,t}$	$r_{s,t}$
ADF Test Statistic	−9.3705	−9.4802	−20.2841
Critical Value (1%)	−3.437	−3.437	−3.437
p-value	0.0000	0.0000	0.0000
Results	Stationarity	Stationarity	Stationarity

Table 3. The ARCH-LM test of three indices logarithmic return series.

	$r_{500,t}$	$r_{K,t}$	$r_{s,t}$
Number of Optimization Iterations	11	10	12
Number of Function Evaluations	59	56	71
Negative Log-Likelihood Function Value	2239.1777	2201.8142	1859.0305
LM Test Statistic	219.0144	273.4263	32.7838
p-value	0.0000	0.0000	0.0003

4 The Construction and Prediction of GARCH Models

4.1 GARCH (1, 1) Model

The GARCH (1,1) model is represented by Eq. (4) as follows:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (4)$$

where σ_t^2 represents the conditional variance at time ‘t’, which denotes volatility. The constant term ω represents the long-term average level of volatility in the model. While ε_{t-1} represents the error term at time ‘t-1’, and the lagged squared coefficient ε_{t-1} reflects the impact of past random shocks on the current volatility. The lagged conditional variance coefficient β_1 reflects the influence of past volatility on the current volatility.

EGARCH (1, 1) Model. The EGARCH (1,1) model is represented by Eq. (5) as follows:

$$\ln(\sigma_t^2) = \omega + g(Z_{t-1}) + \beta_1 \ln(\sigma_{t-1}^2)$$

$$Z_{t-1} = \frac{\varepsilon_{t-1}}{\sigma_{t-1}} g(Z_{t-1}) = \alpha_1 Z_{t-1} + \gamma_1 (|Z_{t-1}| - E(|Z_{t-1}|)) \quad (5)$$

The Exponential GARCH (EGARCH) model, proposed by Nelson in 1990, improves upon the classical GARCH model and captures volatility’s symmetric and asymmetric characteristics more effectively. In the EGARCH model, the coefficient $\gamma_1 < 0$ indicates that negative shocks result in larger changes in volatility compared to positive news.

TGARCH(1,2) Model. The TGARCH(1,1) model is represented by Eq. (6) as follows:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-1}^2 I_{\{\varepsilon_{t-1} < 0\}} + \beta_1 \sigma_{t-1}^2 \quad (6)$$

where $I_{\{\varepsilon_{t-1} < 0\}}$ is an indicator function. When $\varepsilon_{t-1} < 0$, it takes a value of 1, or 0 otherwise. In the TGARCH model, the coefficient $\gamma_1 < 0$ indicates that negative shocks result in larger changes in volatility compared to positive news.

4.2 The Comparative Analysis of GARCH Models

Tables 4, 5 and 6 presents the parameter values of the GARCH (1,1), EGARCH (1,1), and TGARCH (1,1) models for the return time series of the three indices under the Normal Distribution, Student's t Distribution, and GED. Additionally, Table 4 displays the results of the Akaike Information Criterion (AIC). The AIC is used to determine the optimal lag structure and the best-fitting GARCH model, and it is calculated using the following Eq. (7):

$$\text{AIC} = 2k - \ln(L) \quad (7)$$

In the equation, 'k' represents the total number of estimated parameters, and 'L' denotes the likelihood function. Generally, a smaller AIC value indicates better model fit.

Taking into account both the AIC values and the complexity of the models, the mean assumption for all GARCH models in the table is set as AR (30), which assumes that the mean satisfies the following Eq. (8).

$$\mu_t = \varphi_0 + \varphi_1 \mu_{t-1} + \cdots + \varphi_{30} \mu_{t-30} + \varepsilon_t \quad (8)$$

Firstly, comparing the parameter values for each series, we observe that the Student's t distribution exhibits characteristics of peakedness and fat tails, making it more suitable for describing $r_{500,t}$ compared to the Normal distribution. When the thickness or thinness of the tails of the series does not align with the Normal Distribution, GED can adjust the shape parameter 'v' to fit the true distribution form, making it more suitable for describing $r_{K,t}$ and $r_{s,t}$. Then, for each index, the model and distribution corresponding to the minimum AIC value in the sequence distribution will be selected.

Therefore, we select the EGARCH (1,1) model with the Student's t distribution for the S&P 500 return series, the EGARCH (1,1) model with the Normal distribution for the S&P KDBI return series, and the GARCH (1,1) model with the GED distribution for the SSE IFI index return series.

In the following analysis, we will use these selected models to calculate the standard deviation series for the index return VaR calculation, based on the aforementioned selections.

Table 4. Estimation of GARCH (1,1), EGARCH (1,1), and TGARCH (1,1) under three different distributions for the sequence $r_{500,t}$.

$r_{500,t}$ GARCH-AR (30)					
GARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0368	0.1212	0.8518		3013.3990
Students' t	0.0299	0.1238	0.8572		2988.8095
GED	0.0327	0.1169	0.8598		2995.4197
EGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0002	0.0136	0.9895	−0.1378	2986.5782
Students' t	−0.0057	0.0199	0.9882	−0.1718	2955.7619
GED	−0.0030	0.0130	0.9897	−0.1486	2969.6279
TGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0426	0.0662	0.8505	0.0960	3007.7037
Students' t	0.0316	0.0033	0.8865	0.1558	2972.6746
GED	0.0220	0.0000	0.9213	0.1058	2985.6362

Table 5. Estimation of GARCH (1,1), EGARCH (1,1), and TGARCH (1,1) under three different distributions for the sequence $r_{K,t}$.

$r_{K,t}$ GARCH -AR (30)					
GARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0587	0.0633	0.9227		4220.0803
Students' t	0.0667	0.0679	0.9166		4221.2195
GED	0.0625	0.0647	0.9205		4220.6374
EGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0185	0.1009	0.9861	−0.0403	4208.5261
Students' t	0.0219	0.1111	0.9838	−0.0465	4209.3660
GED	0.0193	0.1020	0.9856	−0.0429	4209.5666
TGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0703	0.0280	0.9235	0.0619	4214.7430
Students' t	0.0856	0.0309	0.9116	0.0743	4215.3967
GED	0.0746	0.0281	0.9205	0.0659	4215.4816

4.3 FinTech Risk Measurement and Testing Based on VaR

Risk Measurement. VaR will be calculated using the following formula (9):

$$\text{VaR}_t = \mu_t + \sigma_t \times Z_\alpha$$

(9)

Table 6. Estimation of GARCH (1,1), EGARCH (1,1), and TGARCH (1,1) under three different distributions for the sequence $r_{s,t}$.

$r_{s,t}$ GARCH -AR (30)					
GARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.1733	0.0788	0.8543		3597.4605
Students' t	0.1752	0.0808	0.8538		3577.6959
GED	0.1659	0.0785	0.8582		3574.4294
EGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.0819	0.2080	0.9171	-0.0035	3602.2512
Students' t	0.0832	0.2126	0.9202	-0.0150	3581.2918
GED	0.0764	0.1990	0.9261	-0.0083	3578.6090
TGARCH (1,1)	ω	α_1	β_1	γ_1	AIC
Normal	0.1789	0.0776	0.8503	0.0067	3599.4141
Students' t	0.2048	0.0765	0.8342	0.0289	3579.1936
GED	0.1824	0.0755	0.8466	0.0188	3576.1791

where α is the confidence level and Z_α is a quantile of Standard Normal Distribution under the significance level α . In particular, if α equals 1%, Z_α is 1.6449; if α equals 5%, Z_α is 2.3263. The descriptive statistics of VaR for three index are shown in Table 7:

Table 7. Descriptive statistics of the Value at Risk (VaR)

VaR	S&P 500		S&P KDBI		SSE IFI	
Confidence level	95%	99%	95%	99%	95%	99%
Mean	3.4308	4.8344	8.1206	11.4724	4.4216	6.2670
Standard deviation	6.0733	8.6022	6.3561	8.9831	1.9452	2.7450
Minimum	0.2756	0.4455	1.9772	3.0209	1.9990	2.9529
Median	1.7560	2.4943	6.1840	8.6737	3.7972	5.3410
Maximum	56.9882	80.1332	54.5689	76.4453	13.2896	19.0213

It can be concluded that at the 1% and 5% significance levels, the mean VaR values for the $r_{500,t}$ are 4.83 and 3.43 respectively. This indicates that there is a 99% and 95% probability respectively, for the S&P 500 to control its average loss within 4.83% and 3.43%. Similarly, for the S&P KDBI, there is a 99% and 95% probability, respectively, to control its average loss within 11.95% and 8.46%. For the SSE IFI index, there is a 99% and 95% probability, respectively, to control its average loss within 6.27% and 4.42%. In general, as the confidence level increases, the corresponding VaR values for the index return series also increase, indicating greater volatility in the return series.

Empirical Results. This paper utilizes the GARCH-VaR model to measure the risks in the financial technology sector, using daily return data for the S&P 500, S&P KDBI, and SSE IFI. Empirical results show that: the S&P 500 index has a 99% and 95% probability, respectively, of controlling average losses within 4.83% and 3.43%; The S&P KDBI index has a 99% and 95% probability, respectively, of controlling average losses within 11.95% and 8.46%. The SSE IFI index has a 99% and 95% probability, respectively, of controlling average losses within 6.27% and 4.42%.

Therefore, at the 1% confidence level, we conclude that the mean VaR of the S&P 500 index returns is lower than that of SSE IFI and S&P KDBI. The result indicate that the S&P 500 index can effectively control related risk losses. While compared to the performance of overall financial market, the financial technology sectors both in China and the United States experience higher volatility.

5 Conclusion

In this paper, the empirical results indicate the following: (i) Compared to the overall performance of the financial market, the financial technology sectors in China and the United States exhibit high volatility, which aligns with the characteristics of the technology industry. (ii) Particularly under extreme circumstances, the financial market in the United States, with its stronger innovation capacity and more active trading, faces a higher possibility and magnitude of extreme losses compared to the Chinese financial market. This phenomenon aligns with the high-risk, high-return nature of the financial market.

Moreover, the key to innovation lies in striking a balance between technological advancement and risk control, as well as finding a middle ground between market flexibility and industry regulations. Establishing a regulatory sandbox for financial technology serves as an effective solution for such balance. China and the United States can allow startups to test new financial technology innovations within a defined scope, while regulatory authorities observe and learn from these innovations to inform future regulatory policies.

Finally, in light of the intensified international capital flows and economic globalization, financial markets across countries are tightly interconnected. Consequently, international cooperation in financial markets holds paramount importance. China and the United States should bolster cross-border collaboration and increase data sharing to mitigate extreme event risks. This can be achieved through establishing robust systems for data protection, risk monitoring, and early warning. What is more, regulatory measures can be tailored according to specific risk levels, which can be set by algorithms.

Acknowledgment. 2024 Undergraduate Research Project Funding of Hankou University (Grant Number: 202403).

References

1. Financial stability implications from FinTech. Financial Stability Board (2017). <http://www.fsb.org/2017/06/financial-stability-implications-from-FinTech/>

2. Thakor, A.V.: FinTech and banking: what do we know? *J. Fin. Int.* **41**, 100833 (2020)
3. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**(366a), 427–431 (1979)
4. Giudici, P.: FinTech risk management: A research challenge for artificial intelligence in finance. *Frontiers in Artificial Intelligence* **1**, 1 (2018)
5. Chaudhry, S.M., Ahmed, R., Huynh, T.L.D., Benjasak, C.: Tail risk and systemic risk of finance and technology (FinTech) firms. *Technol. Forecast. Soc. Chang.* **174**, 121191 (2022)
6. Gao, J.: Comparison of FinTech development between China and the United States. *International Journal of Innovative Science and Research Technology (IJISRT)*, www.ijisrt.com. ISSN-2456-**2165**, pp. 1150–1155 (2022). <https://doi.org/10.5281/zenodo.6563524.1272-6>
7. Zhang, H., Bai, C.X.: A Comparative Financial Framework for the Comparison and Insights of Internet Financial Regulation in China and the United States. *Gansu Social Sciences* (2022)
8. Alexander, C., Lazar, E., Stanescu, S.: Forecasting VaR using analytic higher moments for GARCH processes. *Int. Rev. Financ. Anal.* **30**, 36–45 (2013)
9. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007 (1982)
10. Bollerslev, T., Chou, R.Y., Kroner, K.F.: ARCH modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics* **52**(1–2), 5–59 (1992)



A Study of the Time-Varying Shock Effects of Economic Policy Uncertainty on the Prices of Small Agricultural Commodities——Garlic and Ginger as an Example

Jingjing Chen^(✉) and Guixian Tian

School of Finance and Economics, Guangdong University of Science and Technology,
Dongguan, China

chenjingjingxgf@163.com

Abstract. This paper takes garlic and ginger as examples and uses the TVP-SV-VAR model to study the time-varying impact effect of economic policy uncertainty on the prices of small agricultural products in China. The research results show that economic policy uncertainty has dynamic time-varying characteristics on the prices of garlic and ginger, showing a medium- and long-term effect. And due to the characteristics of different small agricultural product categories, there are certain differences in the impact. Secondly, due to the different impacts and degrees of uncertainty events, the impact of economic policy uncertainty on the prices of garlic and ginger at different specific time points also varies.

Keywords: Economic Policy · Uncertainty · Time-Varying Characteristics of Small-Scale Agricultural Products · Economic Fluctuations

1 Introduction

The 20th report stated that it's necessary to speed up building a strong agricultural nation. As agriculture is the foundation of China's economic growth and agricultural products are the mainstay for people's livelihood, the price of such products has long been a national concern. The supply, demand and trading circulation of small agricultural products(SAP) are relatively small compared to large agricultural products, so their price fluctuations are also larger and more frequent, and more susceptible to the impact of uncertain events. Since the outbreak of the financial crisis in 2008, various economic entities have been affected by different degrees of impact; 2020's New Crown Epidemic has also had a huge impact on China's economy, and the price of garlic and ginger has also experienced. The price of garlic and ginger also appeared several large fluctuations, "garlic you hard", "ginger you army" and other phenomena occur frequently. 2008 May garlic price of 1.39 yuan/kg, in October 2010 the price of garlic rose to 12.14 yuan/ kg, an increase of 773%; Ginger's price also rose from 3.49 yuan/kg in May 2008 to 10.07 yuan/kg in November 2010, an increase of 188%.The outbreak of the new crown epidemic in 2020 caused further fluctuations in the prices of garlic and ginger, with the price of

garlic rising from 5.97 yuan/kg in January 2019 to 10.86 yuan/kg in February 2020, an increase of 81.91%; the price of ginger increased from 6.9 yuan/kg in January 2019 to 10.55 yuan/kg in February 2020, an increase of 52.90%. EPU has a greater impact on the prices of garlic and ginger in China. The large fluctuations of SAP cause certain economic losses to farmers, and also bring certain impacts to consumers, and at the same time, it will have certain impacts on the export trade of SAP in China. Based on this, this paper takes garlic and ginger as an example, based on the price fluctuation characteristics of SAP, discusses the impact of economic policy uncertainty (EPU) on garlic and ginger, and on this basis puts forward corresponding countermeasure suggestions, in order to be able to reduce the adverse impact of price fluctuations on production and consumers, which has certain practical significance for the maintenance of price stability of SAP in China.

2 Literature Review

Current academic research on agricultural product prices generally progresses from broad to specific: starting from the price fluctuations of agricultural products in general, then narrowing down to those of SAP, and often analyzing specific SAP as illustrations.

Many scholars explore the underlying causes, with the main factors including supply and demand, costs, and financial aspects. Regarding the supply-demand relationship, Liu Yao employed the SVAR model for an in-depth analysis and concluded that output fluctuations have a relatively minor impact on agricultural product price volatility, while the sales price is more influenced by its own inertia and the money supply. Cost factors have also been studied [1]. Xiao Hao et al. utilized the input-output model to examine the effects of production factors like oil, labor, and land on agricultural product prices, finding that labor has the most significant impact [2]. Yu Wei analyzed from the factor endowment perspective and determined that land has the greatest influence and capital the least. In terms of the international market factor [3]. Chen Aixue believes that fluctuations in international agricultural commodity prices can affect the domestic market [4]. Wang Xiaosong and Xie Shenxiang also share this view, stating that the high integration of domestic and foreign agricultural commodity markets is mainly achieved through international trade. Financial factors are also considered [5].

The price volatility of SAP has drawn the attention of scholars as well. They mainly analyze its volatility characteristics, causes, laws, and influencing factors. Lv Kaiyu and Long Yongzhi investigated the characteristics of small agricultural product price volatility [6]. Zhang Shuzong and Liu Lei, using the cobweb model and VAR model, found that in the long term, the trends of money supply and small agricultural product price increases are consistent, but in the short term, small agricultural product prices are more affected by their own characteristics. Many scholars further analyze the price volatility of specific SAP [7]. For example, Ma Hongyang and Zhao Xia used the ARCH model to empirically analyze the volatility characteristics of garlic, revealing obvious seasonality and cyclicity [8].

In conclusion, existing literature is relatively abundant in research on agricultural product price volatility and the impact of EPU on it. Some studies also use specific agricultural products as examples. However, there are fewer analyses of the price volatility

of SAP due to EPU. Given the differences in supply and demand and market structures of different SAP, this paper takes ginger and garlic as examples and uses the TVP-SV-VAR model to explore the impact of EPU on small agricultural product prices.

3 Theretical Analysis

The determination of the price of SAP is still mainly subject to the supply and demand situation in the market, and according to the theory of equilibrium price and the theory of spider web, when the equilibrium point changes, the equilibrium price will also change, which will lead to price fluctuations. The prices of small agricultural commodities are highly volatile due to their long production cycle. In the market, the supply side will anticipate and make decisions for the current period based on historical conditions, while the demand side will make decisions about demand based on the current price, which ultimately determines the price. In addition to the influence of internal supply and demand, the price of SAP is also more susceptible to external shocks, such as some sudden events or some factors that are difficult to regulate. Therefore, this paper focuses on the impact of EPU on the price of SAP and takes ginger and garlic as examples to analyze.

The impact of EPU on the prices of SAP has two main channels: one is the supply and demand channel. From the supply side, the main manifestation of production costs, production and inventory, when EPU occurs, farmers are very easy to produce negative emotions and expectations, which leads to the volatility of small agricultural commodity prices; from the demand side, mainly contains consumption, investment and export demand. When uncertainty occurs, people are motivated by precaution, leading to a decrease in consumption demand, a decline in investment demand, and a certain degree of impact on export demand, which ultimately affects the price volatility of small agricultural commodities.

Analyzing from the financial channel, when an uncertainty event occurs, the government generally adopts a relatively loose monetary and fiscal policy to stabilize the economy, which leads to a rise in the money supply; in addition, agricultural products are affected by natural factors such as climate, which also leads to a certain amount of speculative capital to enter into that market, which in turn affects the pricing of SAP; in addition, the loose policy will also lead to the anticipation of inflation, and SAP As part of the necessities of life, their prices are bound to be affected to some extent.

EPU mainly through the above two channels ultimately affect the price of SAP, but due to the time lag in the transmission of economic policy, resulting in the shock effect of EPU also exists in a certain time-varying characteristics, which will be verified through empirical methods in this paper.

4 Research Design

4.1 Modeling

A typical SVAR model is shown in Eq. (1) below:

$$Ay_t = F_1y_{t-1} + F_2y_{t-2} + \cdots + F_3y_{t-3} + \mu_t \quad (1)$$

where $t = s+1, \dots, n$, y_t are $k \times 1$ dimensional observation vectors and k is the number of variables. $A, F_1, F_2, \dots, F_s$ are the $k \times k$ order coefficient matrix, μ_t is the structural perturbation matrix, and $\mu_t \sim N(0, \Sigma)$, $\sigma_i (i = 1, \dots, k)$ is the standard deviation of μ_t , then

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_k \end{pmatrix} \quad (2)$$

If the structure shock obeys recursive identification, i.e., the matrix A is lower triangular, the matrix A is denoted as:

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{k1} & \cdots & a_{k,k-1} & 1 \end{pmatrix} \quad (3)$$

Let $B_i = A^{-1}F_i, i = 1, \dots, s$, then Eq. (1) can be further transformed into:

$$y_t = B_1 y_{t-1} + B_2 y_{t-2} + \cdots + B_s y_{t-s} + A^{-1} \Sigma \varepsilon_t \quad (4)$$

where $\varepsilon_t \sim N(0, I_k)$, stacking B_i according to row vectors yields β ($k^2 \times 1$ dimensional vectors), and defining $X_t = I_k \otimes (y_{t-1}', \dots, y_{t-s}')$, where $\otimes$ is the Crocker product, then the *TVP – SV – VAR* model can be shown in Formula 5.

$$y_t = X_t \beta_t + A_t^{-1} \Sigma_t \varepsilon_t \quad (5)$$

If all the parameters in the hypothesis follow a random walk process, the model can capture the gradual or abrupt changes in the economic structure. To avoid the over-identification problem caused by too many parameters, this paper will use the Markov Chain Monte Carlo(MCMC) simulation method for estimation.

4.2 Data Description

This paper mainly selects the data from January 2008 to May 2023 for analysis. The EPU is measured by the index compiled by Scott R. Baker, Nicholas Bloom and Steven J. Davis. This index can effectively measure the level of EPU and has good characteristics of retrospectiveness, continuity and time-variability [9]; and the consumer price index (CPI) is selected as a proxy variable for the economic environment; the prices of SAP are mainly selected to be represented by the national wholesale prices of garlic and ginger. The data of CPI index is from Wind database, and the wholesale prices of garlic and ginger are from Business Forecasting Network. The data of garlic and ginger are weekly data, and this paper uses the mean value to calculate the monthly data.

5 Empirical Analysis

5.1 Data Smoothness Test and Determination of Lag Order

Since the data of each variable are monthly data, in order to avoid the influence of seasonal factors, this paper combed the prices of CPI, ginger and garlic using the CencusX-12 seasonal adjustment method, while taking the natural logarithm for EPU and CPI to reduce the influence of heteroskedasticity. After conducting the stationarity test, it was found that the original sequence is integrated of order one, as shown specifically in Table 1.

Table 1. Results of the stability test

variant	T-value	1% threshold	5% threshold	10% threshold	P-value	Result
LnEPU	−3.0295	−4.0092	−3.4347	−3.1413	0.1271	uneven
LnCPI	−2.7692	−4.0087	−3.4344	−3.1411	0.2107	uneven
GIN	−2.0228	−4.0087	−3.4344	−3.1411	0.5846	uneven
GAR	−2.6752	−4.0089	−3.4345	−3.1412	0.2482	uneven
DLnEPU	−17.3557	−4.0092	−3.4347	−3.1413	0.0000	smoothly
DLnCPI	−6.7985	−4.0122	−3.4361	−3.1421	0.0000	smoothly
DGIN	−8.1708	−4.0092	−3.4347	−3.1413	0.0000	smoothly
DGAR	−3.2773	−4.0089	−3.4345	−3.1412	0.0000	smoothly

Further, the number of lags of the variables was determined, and as shown in Table 2, the finalized minimum order for each variable was determined to be 3, based on the principle of minimization of the information criterion.

Table 2. Diagnostic table of lag order of variables

Lag	LogL	LR	FPE	AIC	SC	HQ
0	−571.4175	NA	0.007831	6.501893	6.573671	6.531003
1	180.1284	1460.632	1.92E-06	−1.809361	−1.450474*	−1.66381
2	202.9901	43.39855	1.78E-06	−1.886894	−1.240898	−1.624903
3	239.7757	68.16755*	1.41e-06*	−2.121759*	−1.188653	−1.743328*
4	248.1356	15.114	1.54E-06	−2.035431	−0.815215	−1.540559
5	254.4283	11.09226	1.72E-06	−1.925744	−0.418419	−1.314431
6	266.1746	20.17435	1.81E-06	−1.877679	−0.083244	−1.149926
7	275.8826	16.23479	1.95E-06	−1.806583	0.274962	−0.962389
8	286.6213	17.47329	2.08E-06	−1.747134	0.62152	−0.7865

5.2 Parameter Estimation Results

In this paper, the initial values of each parameter of the model were set with reference to Nakajima [10], while Monte Carlo simulation was set to iteratively sample 10,000 times and after discarding the first 1,000 samples, the estimation results in Table 3 were obtained, which indicating that the hypothesis that the posterior distributions converge to zero cannot be rejected, and the maximum null factor in the model is 204.17, which means that $10,000/204.17 \approx 49$ uncorrelated samples can be obtained, which is also sufficient for the inference of the posterior distributions, and thus it can be proved that the estimation of the parameters is valid.

Table 3. Parameter estimation results

parameters	average value	(statistics) standard deviation	95% lower bound	95% upper bound	Geweke	ineffective factor
sb1	0.0225	0.0026	0.0183	0.0284	0.013	6.34
sb2	0.0227	0.0027	0.0183	0.0287	0.409	8.18
sa1	0.1061	0.0552	0.0461	0.2507	0.112	46.1
sa2	0.0708	0.0261	0.0388	0.1396	0.692	72.1
sh1	0.3701	0.1016	0.1944	0.5949	0.706	130.57
sh2	0.668	0.1311	0.4385	0.9495	0.039	204.17

5.3 Characterization of Time-Varying Parameters

The first three rows of Fig. 1 represent the sample autocorrelation coefficients, sample paths and density functions calculated by the MCMC sampling technique, respectively. From the sample autocorrelation coefficients in Fig. 1, the autocorrelation coefficients of S_b , S_a , and S_h all decrease to fluctuate around 0, which indicates that the samples are not autocorrelated; however, the convergence of the first two parameters is faster, while the convergence of S_h is a bit slower, which indicates that the MCMC method is effective for the simulation of parameter distributions. In terms of the sample paths, the paths of the samples are basically smoother, indicating that the MCMC method of taking values can produce effective samples; the density function also further indicates that the posterior distribution is more concentrated, thus indicating that the model is effective.

5.4 Equal Time Interval Impulse Response Results

The TVP-SV-VAR model can simulate impulse responses across different time periods. In this study, lags of 4, 8, and 12 are chosen to signify the short, medium, and long term respectively, for the purpose of analyzing the time - varying correlations among the variables.

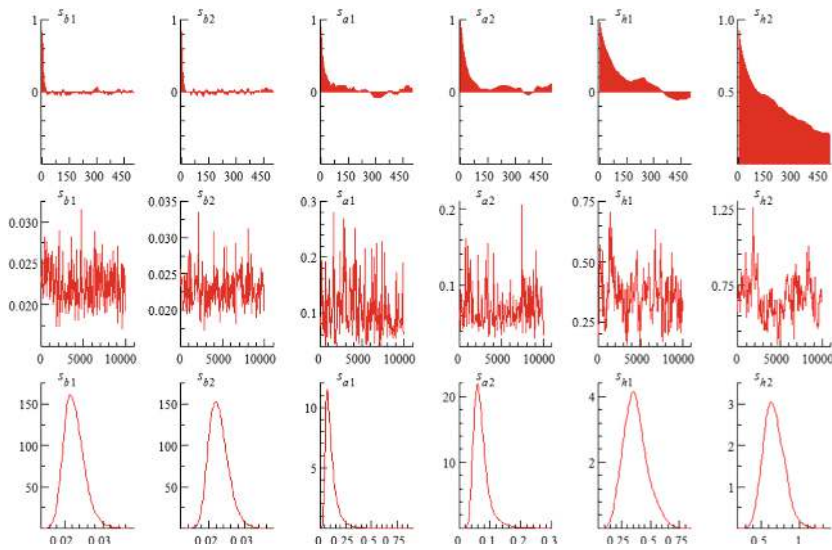


Fig. 1. Sample autocorrelation coefficient, sample path and density function

In general, the trend of equal-interval impulse responses of each variable under different interval periods is roughly similar, with the impulse response curve of lag 4 being relatively smooth, and the impulse response curves of lag 8 and lag 12 having larger fluctuations, which implies that, due to a variety of factors, the prices of SAP, such as ginger and garlic, have sufficient time to respond to the shocks of EPU.

Figure 2 shows the impact of EPU on inflation, it can be seen from the figure, the impact of EPU on China's consumer price index before 2019 are fluctuating around 0, and “positive-negative-positive” impact. Small peaks appeared in 2008, around 2011, around 2015, which is related to the financial crisis, the European debt crisis and the domestic stock market crash has a certain impact; 2019–2020 to reach the peak and show a positive impact, which is closely related to the impact of the new crown epidemic. This suggests that with EPU events can have an impact on the domestic economy, which in turn has an impact on the prices of small agricultural commodities.

From Fig. 3, the impacts of EPU on ginger and garlic are different, but the short-, medium- and long-term trends are the same. Before 2018, the impact of EPU on ginger is mainly negative, and after 2018 it starts to show a positive impact and reaches its maximum in 2020. The impact on garlic is mainly characterized by the cycle of “negative-positive-negative”, which is mainly a negative impact before 2011 and turns into a positive impact after 2012; it starts to show a negative impact after 2014 and reaches a low point in 2015, and then turns into a positive impact and reaches a peak after 2019. Considering that ginger and garlic, as SAP, have certain production cycles, coupled with the uncertainty of economic policies that can have different impacts on market supply and demand, the shocks appear with a certain lag. Specifically, the uncertainty of economic policy leads to changes in the domestic and international environment, which in turn has a certain impact on the agricultural market and affects the expectations of both supply

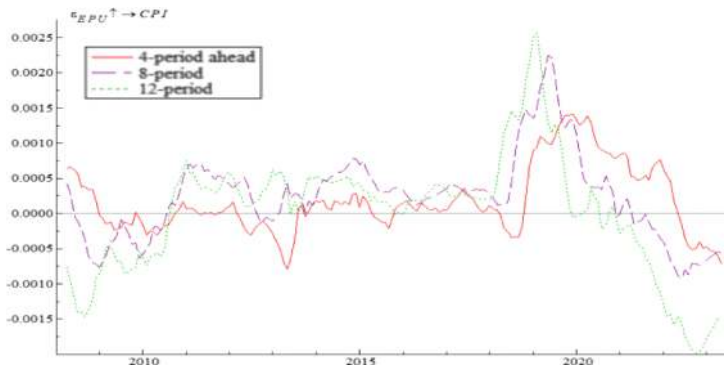


Fig. 2. Equally spaced impulse response of economic uncertainty to CPI

and demand, thus causing a certain impact on the price of ginger and garlic. In general, EPU presents dynamic time-varying characteristics for ginger and garlic.

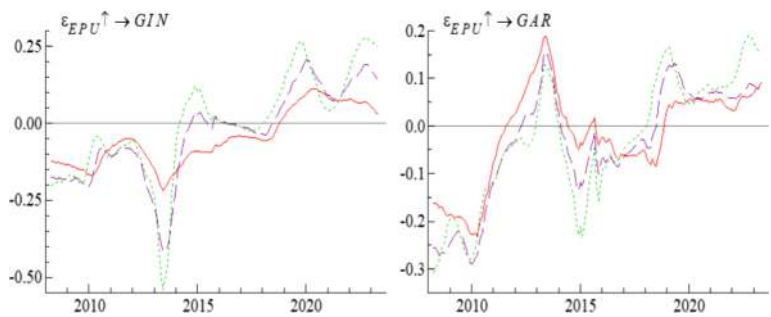


Fig. 3. EPU for ginger and EPU for garlic

5.5 Characterizing the Time-Point Impulse Response

Referring to Qing Yuxi et al. [11], this paper selects August 2008, October 2011, June 2018 and January 2020 as the representative time points for the four major emergencies: the financial crisis, the European debt crisis, China-US trade friction, and the new crown epidemic.

Figure 4 shows the time-point-specific impulse response of EPU on the consumer price index. From the figure, it can be seen that the financial crisis and the new crown epidemic period, the impact of EPU on the consumer price index is relatively similar, when the EPU occurs, it will cause a strong impact on the CPI, followed by a gradual weakening of the impact, which suggests that the financial crisis and the new crown epidemic caused by the liquidity crisis and the panic situation of the population directly on the market of SAP caused a strong impact, with policy With the adjustment of policies and the easing of the crisis and epidemic, the prices of SAP reside in decline; the impact

on the CPI during the European debt crisis is generally stable; during the China-US trade friction, the influence of EPU on CPI gradually increases. This indicates that the trade friction affects consumers' consumption expectations and causes greater price fluctuations.

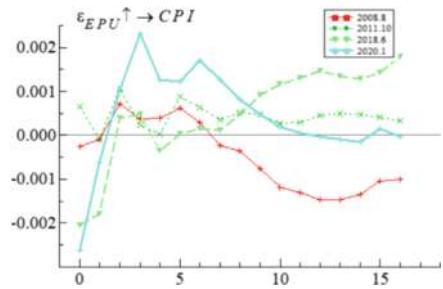


Fig. 4. Time-point-specific impulse responses of EPU to the CPI

Figure 5 shows the trend of the impact of EPU on ginger and garlic at the same specific point in time is roughly similar. Among them, during the period of Sino-US trade friction and the new crown epidemic, the impact of EPU on ginger and garlic shifted from a negative shock to a positive shock, which is due to the fact that when the unexpected event occurs, the people's expectation of the event occurs for a short period of time, and as the front of the unexpected event lengthens, the people form a certain expectation, which in turn reacts in the price, but as the new crown epidemic stabilizes, the price of the small prices of agricultural products also gradually stabilized. During the European debt and financial crises, the impact of EPU on ginger and garlic was mainly negative and then began to converge towards zero.

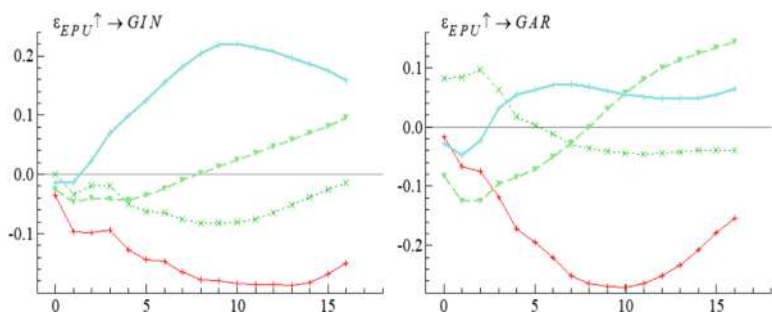


Fig. 5. EPU for ginger and EPU for garlic

6 Conclusion

This paper analyzes the dynamic shocks of EPU on garlic and ginger using the TVP-SV-VAR model and analyzes the movement of CPI under EPU and the shock effect of the CPI on garlic and ginger. The study finds: first, the impacts of EPU on garlic and ginger have obvious time-varying characteristics, and the impacts mainly reflect medium-to long-term effects. Second, the impact of EPU on CPI mainly shows a “positive-negative-positive” cyclical fluctuation, but the impact is not significant, basically fluctuating around 0, but in the period of the New Crown Epidemic, the impact is mainly positive. Third, the impact of CPI on garlic and ginger prices is different. It is mainly a negative shock for garlic and a positive shock for ginger, due to differences in the characteristics of the different SAP. However, the impact of CPI on garlic and ginger is broadly similar at different specific points in time.

Acknowledgment. Guangdong University of Science and Technology Research Projects: Research on the Impact of Digital Economy on the Optimization and Upgrading of the Industrial Structure of Dongguan’s Manufacturing Industry(Project No.GKY-2022KYYBW-2).

References

1. Yao, L.: Research on price volatility mechanism of agricultural products in China-Based on SVAR model. *Shandong Social Science* **04**, 146–151 (2017)
2. Hao, X.I.A.O., Shu, L.I.U., Cuihong, Y.A.N.G.: Analysis of supply factors of agricultural price increase:based on the perspective of cost transmission capacity. *Agricultural Technology and Economics* **06**, 80–91 (2014)
3. Yu, W.: Research on the impact of agricultural factor endowment on the price volatility of agricultural products. *Price Theory and Practice* (12), 82–85+200 (2021)
4. Aixue, C.: The impact of international agricultural price fluctuations on China and counter-measures. *Economic Zongheng* **02**, 102–107 (2016)
5. Xiaosong, W., Shenxiang, X.: How have international agricultural commodity prices affected Chinese agricultural commodity prices? *Econ. Res.* **47**(03), 141–153 (2012)
6. Kaiyu, L., Yongzhi, L.: Characterization of price volatility of small agricultural commodities. *World Agriculture* **12**, 51–53 (2011)
7. Zhang, S., Liu, L.: Price volatility of small agricultural commodities and the impact of money supply--a mathematical and empirical study based on spider web model and VAR model. *Financial Theory and Practice* (11), 75–79 (2012)
8. Ma, H., Zhao, X.: Empirical analysis of price volatility characteristics of small agricultural products in China-Taking garlic as an example. *Agricultural Technology and Economics* **06**, 33–48 (2021)
9. Nan, Y., Hou, T., Song, M.: Dynamic time-varying effects of economic policy uncertainty on urban residents’ consumption behavior—an empirical test based on TVP-SV-VAR model. *Contemporary Economic Research* **01**, 118–128 (2022)
10. Nakajima, J.: Time-varying parameter VAR model with stochastic volatility: an overview of methodology and empirical applications. *Monetary and Economic Studies* **29**(6), 107–142 (2011)
11. Yuxi, Q., Fenglong, G.: Economic Policy Uncertainty, International Crude Oil Price Changes and Futures Prices of Agricultural Products in China-Analysis Based on TVP-SV-VAR Model. *Financial Economy* **12**, 79–93 (2022)

Business Intelligence and Decision Support



Research on Key Technologies and Innovative Strategies for Precision Marketing of Special Agricultural Product E-commerce Based on User Profile

Zhenyan Hu and Qiongjie Zhou^(✉)

School of Management, Wuhan Donghu College, Wuhan, China
zqj_dhxy@163.com

Abstract. With the development and popularization of the Internet, the e-commerce platform for featured agricultural products has become one of the important channels for agricultural product marketing. However, achieving precise marketing of characteristic agricultural products remains a challenge. This article explores the application of data mining and machine learning algorithms in user behavior analysis and prediction in precision marketing, as well as the design and optimization of personalized recommendation systems, through the construction and analysis of user profiles. It proposes innovative marketing strategies such as precise positioning and customized marketing plans based on user profiles, and cooperative marketing models in cooperation with agricultural production enterprises. By studying the key technologies and innovative strategies of precision marketing in specialty agricultural product e-commerce, more effective marketing solutions can be provided for specialty agricultural product e-commerce platforms, which can effectively promote the development of the agricultural product industry.

Keywords: User Profile · Agricultural Product E-commerce · Precision Marketing · Data Mining · Personalized Recommendations

1 Introduction

With the rapid development of Internet technology, the wave of e-commerce is rising rapidly. The e-commerce platform for featured agricultural products has become one of the important channels of agricultural product marketing [1]. The origin of agricultural products has unique regional and cultural characteristics, and the traditional sales model has been unable to meet consumers' demand for quality and service. Therefore, the sales of featured agricultural products are often subject to the limited marketing channels and asymmetric sales information. User profiling, as an abstract model for describing user characteristics and behaviors, is an important foundation of precision marketing. By constructing and analyzing user profiles, we can better understand user needs and behavioral intentions, and provide personalized and customized services for characteristic agricultural product e-commerce platforms. Data mining and machine learning

algorithms also play a significant role in user behavior analysis and prediction, providing strong support for marketing activities by mining the patterns and trends behind data. The design and optimization of personalized recommendation systems are also important means to improve user stickiness and purchase rates. By analyzing user preferences and behavioral history, it is possible to recommend characteristic agricultural products that meet their interests and needs, improving user experience and loyalty. On the basis of precision marketing, innovative marketing strategies are also key to achieving the development of characteristic agricultural product e-commerce. Based on user profiling, precise positioning and customized marketing plans can better allocate resources and energy to the truly needed user group, effectively improving marketing efficiency and effectiveness [2]. The cooperative marketing model with agricultural product production enterprises is also an important innovative direction. Through deep cooperation with agricultural product production enterprises, risks and benefits are shared, production and sales are connected, resources are integrated, and the coordinated development of the entire agricultural product industry chain is promoted. The research on precision marketing technology and innovative strategies for characteristic agricultural product e-commerce based on user profiling is of great significance. Therefore, this article explores how to develop innovative precision marketing strategies through key technological means such as user profiling analysis, data mining, machine learning, and personalized recommendations, providing more effective marketing solutions for characteristic agricultural product e-commerce platforms, and promoting the digital transformation and intelligent development of the agricultural product industry.

2 Overview of Precision Marketing Theory and User Profile Construction Technology

2.1 Definition and Characteristics of Precision Marketing

Precision marketing is a personalized marketing strategy aimed at specific user groups. By deeply understanding user needs and behaviors, it provides customized products and services to improve user satisfaction and marketing effectiveness. It focuses on deep understanding and mining of user needs. By analyzing user behavior data and preferences, it can understand user consumption habits, purchasing motivations, and preferences, and provide personalized recommendations and customized services that meet their needs. Precision marketing emphasizes personalized recommendations and customized services. Based on the characteristics and behaviors of users, it utilizes recommendation algorithms and technical means to recommend characteristic agricultural products that meet their preferences and needs, thereby improving user purchase rates and loyalty. By establishing user profiles and maintaining a user information database, precise positioning and personalized services for users can be achieved, stable and lasting user relationships can be established, and user participation and loyalty can be promoted. Precision marketing relies on data-driven and technological support, and requires the use of techniques such as data mining, machine learning, and artificial intelligence to analyze and mine user data in order to achieve personalized recommendations. The key to precision marketing lies in accurately constructing user profiles. User profiles are

abstract descriptions of user characteristics and behaviors, including important information such as basic information, interests, and consumption habits [3]. Through data collection, cleaning, analysis, and modeling, real and comprehensive user profiles can be constructed to achieve the goal of precision marketing. When collecting data, obtain user behavior data, transaction records, and social interaction information through multiple channels to establish a multi-dimensional consumption model for users. In the data cleaning stage, data is deduplicated, filtered, and organized to ensure accuracy and credibility. The data analysis steps apply techniques such as data mining and statistical analysis to reveal user behavior patterns and preference trends, providing support for establishing user patterns. In the data modeling stage, machine learning and deep learning are used to construct user patterns, achieving accurate description and prediction of user information. Through in-depth research and application of precision marketing theory and user model construction technology, we aim to provide more effective marketing solutions for characteristic agricultural product e-commerce platforms and promote the development and innovation of the agricultural product e-commerce field.

2.2 Application of Precision Marketing in the E-commerce Field

In order to better meet consumer needs, improve marketing effectiveness and user experience, specialty agricultural product e-commerce platforms should adopt precision marketing. As a special form of e-commerce, this platform needs to rely on precision marketing theory to achieve more effective sales and promotion. By accurately targeting the target user group and implementing precise push and marketing based on user geography, preferences, and other characteristics. Personalized recommendations and customized services can enhance the user experience of conveniently purchasing specialty agricultural products, increase purchase rates and satisfaction. Utilizing user profile construction technology to deeply analyze user consumption behavior and preferences, providing the platform with more accurate user profiles. Optimizing and updating user profiles can achieve precise positioning and customized marketing, enhancing user stickiness and loyalty. Big data technology provides insights into user needs, supports marketing activities for specialty agricultural products, and enhances marketing effectiveness and revenue. The architecture of the precision marketing platform for characteristic agricultural products includes four parts: data collection, preprocessing, analysis, and decision support, as shown in Fig. 1. Precision marketing helps characteristic agricultural product e-commerce platforms better understand users and their needs, provide personalized services and recommendations, enhance user experience and purchase intention. Through precision marketing theory and user profile construction technology, characteristic agricultural product e-commerce platforms can achieve more precise and efficient marketing methods, promote sales and promotion, and promote digital transformation and intelligent development [4].

2.3 Concept and Construction Method of User Profile

In the field of specialty agricultural product e-commerce, implementing precision marketing is of great significance for improving product sales and expanding the market. User profile is a user feature description constructed based on user data, behavior, and

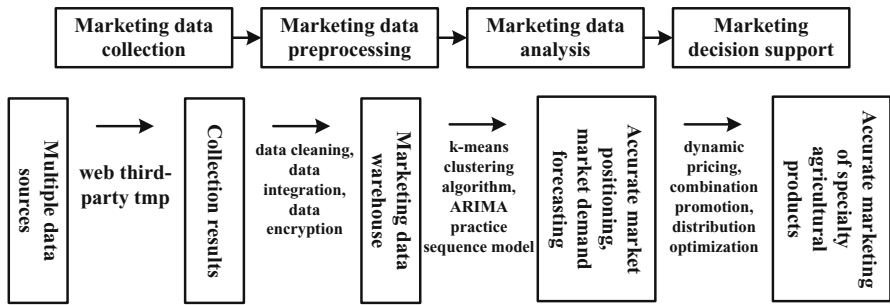


Fig. 1. Architecture of precision marketing platform for special agricultural products.

preferences, aiming to more accurately understand user needs, preferences, and purchasing motivations. User profiles not only help companies better understand the target user group, but also provide data support and guidance for precision marketing. User profile refers to the comprehensive description and analysis of user characteristics and behaviors through the collection, analysis, and integration of multidimensional data such as personal information, consumption behavior, and social interaction. The purpose of profiling consumers is to better understand them and achieve precise personalized marketing. For the construction of user profiles, the first step is to collect data, which can come from multiple channels such as user consumption records, browsing behavior, and online interaction. At the same time, it is necessary to properly clean, integrate, and summarize the data to ensure its accuracy and reliability [5]. Next, it is necessary to use techniques such as data mining and machine learning to analyze the collected data, discover user behavior characteristics and consumption habits, and provide a basis for the construction of user profiles. In the data modeling process, clustering analysis, association rule mining, and other methods can be used to establish user profile models, classify and describe users. The construction of user profiles also needs to consider multidimensional features. In addition to the basic information of users, it is also necessary to consider information such as their interests, social relationships, purchasing preferences, etc., in order to more comprehensively describe user characteristics. As user behavior and needs change, user profiles also need to be updated and adjusted in real-time to maintain accurate descriptions and effective analysis of users. The steps to build a user profile model are shown in Fig. 2. In the field of specialty agricultural product e-commerce, user profile construction technology plays a crucial role. By establishing accurate and comprehensive user profiles, enterprises can gain a deeper understanding of user needs and behaviors, achieve personalized marketing and precise recommendations, and thereby improve user experience and sales effectiveness.

2.4 The Role of User Profiles in E-commerce Marketing

In the field of specialty agricultural product e-commerce, implementing precision marketing requires fully utilizing user profile construction technology to ensure effective positioning and meeting user needs. By collecting and analyzing user personal information, consumption preferences, purchasing behavior, and other data, establishing a

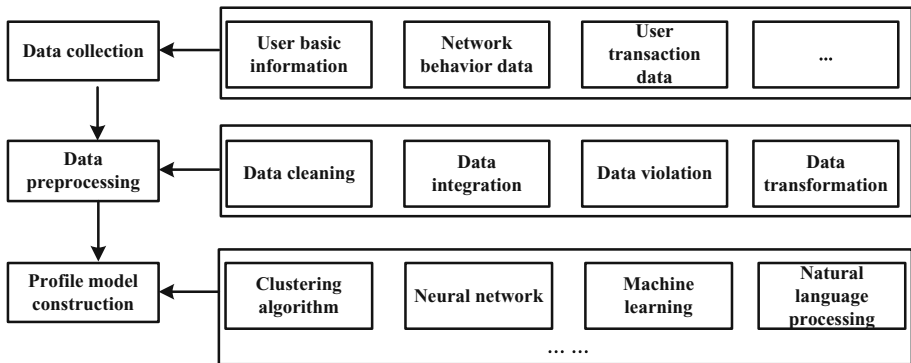


Fig. 2. Construction of user profile.

complete user profile can help enterprises more accurately understand the needs and preferences of target users. This can provide enterprises with more precise market positioning and marketing strategy formulation, while achieving personalized recommendations and customized services [6]. Enterprises can provide personalized product recommendations and customized service solutions based on user preferences and behavioral habits, thereby improving the purchasing experience and satisfaction of users. By finely analyzing user profiles, enterprises can more accurately choose promotion channels and advertising content to improve ad click through rates and conversion rates, and achieve more effective marketing effects. Continuously updating and optimizing user profiles helps companies establish long-term stable user relationships, optimize products and services, enhance user loyalty and satisfaction, and build a good user reputation and brand image. In the e-commerce of characteristic agricultural products, by deeply analyzing user profile data, enterprises can better understand user needs and behaviors, provide personalized products and services, and achieve the goal of precision marketing.

3 Key Technologies for Precision Marketing of Special Agricultural Product E-commerce Based on User Profile

3.1 Data Collection and Processing

Data plays an important role in the process of building user profiles. By collecting and processing user data, it is possible to gain a deeper understanding of user characteristics, needs, and behaviors, providing data support for precision marketing. Data collection is the first step in building user profiles and can be achieved through various methods, including user registration information, purchase records, browsing behavior, social interaction, etc. Through this data, key information such as user basic information, preferences, and purchasing habits can be obtained, laying the foundation for building accurate user profiles [7]. Data cleaning and integration is an important part of data processing, which involves a large amount of data in the data collection process. It is necessary to clean and integrate the data, remove duplicate and incorrect data, ensure data accuracy and integrity, and solve inconsistent format problems for subsequent analysis

and mining. Data analysis is the core link of data processing, which uses data analysis tools and algorithms to deeply analyze the cleaned data, excavate user behavior characteristics and consumption preferences, provide a basis for user profile construction, help discover user group characteristics, better understand user needs and behavior patterns, and achieve precision marketing. Data updating and optimization is a continuous work of data. As user behavior and needs change, user profiles need to be constantly updated and optimized. Regular analysis of user data, adjustment and improvement of user profile models, and maintaining accuracy and effectiveness, provide continuous data support for precision marketing. Data collection and processing are important technical links in the precise marketing of specialty agricultural product e-commerce. Standardizing and accurately collecting and processing user data, tapping into the potential of data, better understanding user needs, achieving personalized marketing and precise recommendations, improving user satisfaction and marketing effectiveness, and promoting the healthy development of specialty agricultural product e-commerce. Effective data collection and processing technology is one of the key factors in precision marketing of specialty agricultural product e-commerce.

3.2 User Profile Construction Technology

The dimensions and indicator system of user profile in user profile construction technology, as well as the application of data analysis and mining technology in user profile construction, are two crucial aspects. In characteristic agricultural product e-commerce, the dimensions of user profile can include basic personal information, consumption habits, user behavior, social relationships, and other aspects, as shown in Fig. 3. By establishing a comprehensive and accurate user profile indicator system, the characteristics and needs of users can be described and expressed more comprehensively and accurately, providing a basis for subsequent precision marketing. User behavior data, transaction data, evaluation and other information can be collected through data collection tools, and data mining techniques can be used to analyze and process the data. Through statistical analysis, feature extraction, attribute association and other methods of user data, user consumption preferences, behavior patterns, potential needs and other information can be mined. Meanwhile, through machine learning algorithms, users are grouped and clustered to discover similarities and differences between user groups, providing a foundation for personalized recommendations for precision marketing. User profile construction technology is an indispensable part of precision marketing in specialty agricultural product e-commerce. By establishing a user profile system with rich dimensions and clear indicators, and using data analysis and mining technology, it can more accurately grasp user needs, achieve personalized recommendations and precision marketing [8]. Special agricultural product e-commerce platforms should focus on establishing and optimizing user profiles, continuously improving the accuracy and precision of user profiles, thereby improving user experience, enhancing user loyalty, and promoting sales performance growth.

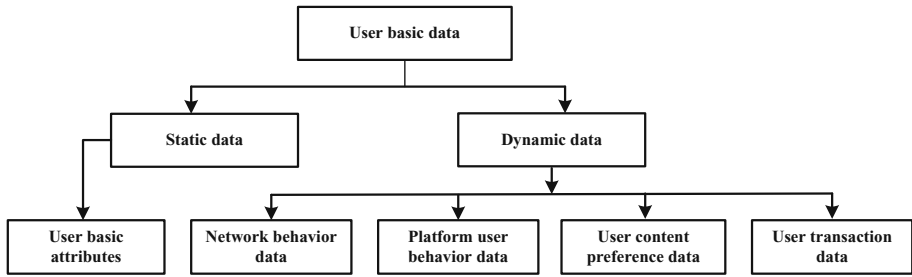


Fig. 3. User basic data structure diagram.

3.3 Development of Precision Marketing Strategies

In the e-commerce of characteristic agricultural products, the formulation of precise marketing strategies is the key to achieving marketing goals and improving user satisfaction. Through user profiling based technology, specialty agricultural product e-commerce can develop more effective and precise marketing strategies to meet user needs and increase sales. The establishment and analysis of user profiles are the basis for formulating precise marketing strategies. The characteristic agricultural product e-commerce platform collects user personal information, browsing records, transaction behavior and other data to establish a complete user profile and deeply explore user preferences, interests, and purchasing habits. Data analysis and mining techniques can conduct detailed analysis to determine the common characteristics and needs of user groups, providing data support for precision marketing strategies. Developing personalized marketing strategies based on user profiles can effectively improve marketing effectiveness. Specialized agricultural product e-commerce can design customized marketing strategies based on user profiling results, promote personalized product recommendations and promotional activities [9]. Personalized strategies can attract user attention, accurately recommend to relevant consumers who are interested in related products, meet consumer purchasing needs, and improve conversion rates and user loyalty. This characteristic agricultural product e-commerce platform also needs to monitor marketing effectiveness, adjust and optimize strategies based on data analysis results, continuously attract users and promote sales. The precise marketing logic system for agricultural products based on user profiling technology is shown in Fig. 4. With the improvement of data technology and intelligence level, specialty agricultural product e-commerce will have more opportunities to utilize user profiling technology to innovate marketing strategies, provide better user experience, and achieve business goals.

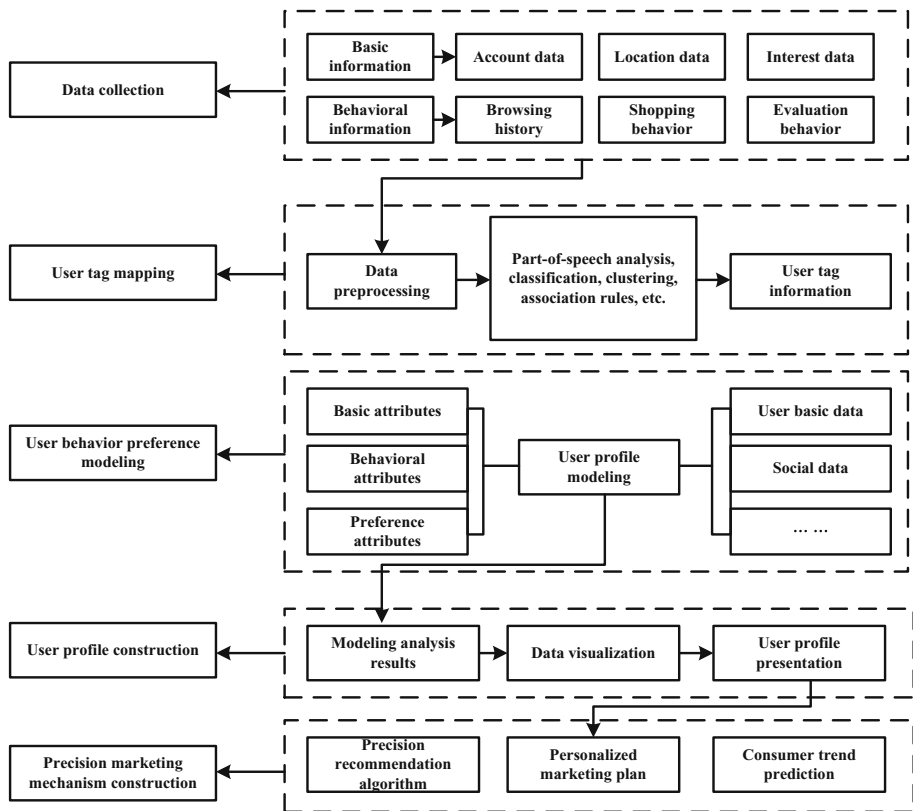


Fig. 4. Logical system of precision marketing for agricultural products based on user profile.

4 Innovative Strategies for Precise Marketing of 4 Characteristic Agricultural Product e-commerce Platforms

4.1 Product Strategy Innovation

In terms of product strategy innovation, specialty agricultural product e-commerce can meet user needs, enhance user experience, and promote sales growth by optimizing product features, quality standards, and product combinations. The innovation of product features is an important means for implementing precision marketing in specialty agricultural product e-commerce. Special agricultural products have unique regional cultures, production processes, and taste characteristics. Sellers stationed on agricultural e-commerce platforms can enhance these characteristics, create a unique product image, and improve product recognition and attractiveness [10]. The innovation of quality standards is also an important direction for product strategy innovation. E-commerce of characteristic agricultural products should pay attention to the quality and safety assurance of products, strictly control the product supply chain, and ensure that the source of products is traceable and the quality is guaranteed. Specialty agricultural product

e-commerce can design diverse product combinations based on user profiles and consumption habits to meet the needs of different users. By flexibly combining packages and limited time promotions, specialty agricultural product e-commerce can increase product sales and user purchasing desire, achieving sales growth and market share enhancement. By strengthening product features, improving quality standards, and diversifying product combinations, specialty agricultural product e-commerce can better meet user needs, create competitive product advantages, promote sales growth, and enhance user loyalty. In the future, specialty agricultural product e-commerce should continue to pay attention to market changes and user needs, continuously optimize product strategy innovation, maintain the competitive advantage of specialty agricultural product platform holders, and achieve sustainable social development.

4.2 Price Strategy Innovation

By adjusting pricing policies, differentiated pricing, and flexible promotions, specialty agricultural product e-commerce can better meet the needs of different user groups, improve the attractiveness and competitiveness of products. The reform of price policy is the core of price strategy innovation. E-commerce of specialty agricultural products can incentivize users to purchase and increase sales through diversified pricing policies (such as discounts, discounts, buy giveaways, and other promotional activities). Based on market demand and product characteristics, specialty agricultural product e-commerce can also adopt differential pricing strategies to provide products at different prices to different user groups, expand customer base, and increase market share. Differentiated pricing is an important means of price strategy innovation. Specialty agricultural product e-commerce can develop differentiated pricing strategies based on different product characteristics and user needs. Through differentiation in product quality, brand image, service level, and other aspects, it can assign different values to product pricing, attract different consumer groups, and improve market acceptance of prices. Through differentiated pricing, specialty agricultural product e-commerce can consolidate market position and improve competitiveness. Flexible promotion strategies are also an important aspect of price strategy innovation. Special agricultural product e-commerce can combine holidays, special events, and other opportunities to launch various promotional activities, such as limited time discounts, buy one get one free, etc., to stimulate user purchasing desire and increase product sales. Special agricultural product e-commerce can also attract users to repeat purchases and improve user loyalty through membership systems, point deductions, and other means [11]. Implementing price strategy innovation in precision marketing for specialty agricultural product e-commerce is an important means to achieve commercial success. Through price policy adjustments, differentiated pricing, and flexible promotions, specialty agricultural product e-commerce can better meet user needs, stimulate sales growth, improve market share and user loyalty. In the future, specialty agricultural product e-commerce should continuously optimize pricing strategies and innovate based on market changes and user feedback, maintain market competitive advantages, and achieve long-term development.

4.3 Channel Strategy Innovation

By optimizing channel layout, integrating multiple channels, and developing emerging channels, characteristic agricultural product e-commerce can expand sales channels, improve user experience, increase sales revenue, and achieve precise marketing goals. By analyzing user profiles and consumer behavior, specialty agricultural product e-commerce can identify target user groups and their purchasing habits, and select suitable online and offline sales channels. On the online side, specialty agricultural product e-commerce can be promoted and sold through social media, e-commerce platforms, etc.; Offline, you can choose logistics distribution centers, agricultural product display stores, etc. for sales display. Optimizing channel layout can better cover user groups, increase product exposure and sales rate. Multi channel integration achieves unified management of product information, inventory, and logistics through the integration of online and offline channels, providing a more convenient and efficient shopping experience. Developing emerging channels is also an important means of channel strategy innovation. Specialty agricultural product e-commerce can explore new sales channels through partnership relationships, such as cooperative farms, community group buying, etc., to expand the sales market. Special agricultural product e-commerce can also explore new technology applications, such as unmanned stores, intelligent express boxes, etc., innovate marketing channels, improve shopping experience and user stickiness. Implementing channel strategy innovation in precision marketing of specialty agricultural product e-commerce is a key strategy to enhance competitiveness and achieve growth. Through the above methods, specialty agricultural product e-commerce can better serve users, improve sales efficiency and user satisfaction, and achieve business goals. In the future, specialty agricultural product e-commerce should continuously pay attention to market demand and trends, flexibly adjust channel strategies, maintain market sensitivity, and achieve continuous innovation and development.

4.4 Innovation of Promotional Strategies

By optimizing promotion methods, personalized promotion, and interactive promotion, specialty agricultural product e-commerce can attract more users, improve user experience, increase transaction frequency, and achieve the goal of precision marketing. Optimizing promotional methods is the key to innovative promotional strategies. E-commerce for specialty agricultural products can adopt various promotional methods such as discount promotions, full discounts, and limited time discounts to make consumers feel the joy of affordability and shopping, thereby promoting sales growth, strengthening user awareness and trust in the product, and promoting user participation and the formation of purchasing decisions. Personalized promotion is a key direction for innovative promotion strategies. E-commerce of specialty agricultural products can tailor personalized promotion plans based on user profiles and purchase history, providing users with customized promotion information and discounts, stimulating their participation and purchasing desire, further meeting user needs, improving user satisfaction, and promoting user loyalty. Interactive promotion is also an important way to innovate promotional strategies. Specialized agricultural product e-commerce can enhance user engagement and interactivity by organizing online and offline interactive

marketing activities (such as lucky draws, online live broadcasts, etc.), attract user attention, increase user interaction frequency and feedback effect, enhance user stickiness and word-of-mouth dissemination effect. Implementing innovative promotion strategies in precision marketing for specialty agricultural product e-commerce is an important strategy for achieving market promotion and sales growth. By optimizing promotion methods, personalized promotions, and interactive promotions, specialty agricultural product e-commerce can attract more users, increase transaction times, promote sales and user satisfaction. In the future, specialty agricultural product e-commerce should keep up with market development, continuously optimize promotion strategies, enhance market competitiveness, and achieve long-term sustainable development.

5 Conclusion

Through methods such as user profiling analysis, precise positioning of user needs, and implementation of personalized marketing, specialty agricultural product e-commerce can better meet user needs, enhance user experience, increase sales, and achieve precision marketing goals. This study proposes a series of suggestions and insights in exploring the key technologies and innovative strategies of precision marketing for specialty agricultural product e-commerce. In the development of characteristic agricultural product e-commerce, analyzing user profiles and accurately positioning user needs is crucial. The construction and application of user profiles can help e-commerce more accurately understand user needs and implement personalized marketing strategies. Subsequently, key technologies of precision marketing were discussed, such as data mining, artificial intelligence, big data analysis, etc., to provide technical support for e-commerce, better analyze user data, and mine user needs. Finally, innovative marketing strategies were introduced, such as price strategy, channel strategy, promotion strategy, etc., to attract more users, increase sales, and achieve precise marketing goals. For the development of characteristic agricultural product e-commerce, research on key technologies and innovation strategies is crucial, and future research directions and suggestions are proposed. E-commerce should optimize technology and strategies based on actual situations, enhance competitiveness, and meet user needs. This research is of great significance for improving enterprise competitiveness, market share, and user satisfaction. In the future, it should continue to conduct in-depth research, explore innovation, and promote the application of precision marketing strategies in the characteristic agricultural product industry to achieve better commercial benefits and user experience.

Acknowledgment. This work was supported by the grants from Youth Foudation Wuhan Donghu University in 2023 “Research on Network Marketing Strategies for Hubei Province’s Special Agricultural Products from the Perspective of All Media” (Project No.: 2023dhsk005).

References

1. Jinyan, W.: The application of big data technology in precision marketing of special agricultural products. *China New Technology and New Products* **20**(5), 137–139 (2024)

2. Ya, W.: Exploration of building a data driven precision marketing model. *Old Brand Marketing* **24**(5), 13–15 (2024)
3. Xia, H.: Analysis and research on precision marketing strategies in E-commerce under the background of big data. *Old Brand Marketing* **18**(4), 45–47 (2024)
4. Jia, T.: Research on precision marketing model of E-commerce platforms based on data mining technology. *Modern Business* **23**(24), 11–14 (2023)
5. Wenqi, Z., Xu, Y.: Precision marketing management and strategy exploration of Songshan Shouwu tea E-commerce under the background of big data. *Food Research and Development* **44**(21), 229–230 (2023)
6. Ziyun, Q.: Research on precision marketing methods for E-commerce platforms based on cluster analysis. *Digital Communication World* **11**(10), 188–190 (2023)
7. He, X., Gao, X., Lin, Y.: Research on precision marketing strategies for sports goods based on user profile. *Northern Economic and Trade Journal* **46**(10), 65–70 (2023)
8. Ge, R.: Research on precision marketing strategies of commercial banks driven by financial customer portraits. *Shanghai Commercial* **18**(09), 39–41 (2023)
9. Heli: Precision Marketing Management and Strategy Exploration of Longquanyi Peach E-commerce under the Background of Big Data. *Food Research and Development* **44**(09), 231–232 (2023)
10. Huang, L., Chen, Y.: Design and implementation of agricultural product E-commerce platform based on user profile. *Computer Programming Skills and Maintenance* **21**(02), 34–36 (2021)
11. Huang, J., Yin, Z., Lin, X., et al.: User profile analysis of agricultural products based on network data: a case study of Xuzhou burdock sauce. *Business News* **31**(23), 7+9 (2020)



Research on Intelligent Financial Risk Identification and Early Warning Model Driven by BDP

Xiaodi Fan and Ge Guo (✉)

School of Management, Wuhan Donghu College, Wuhan, China

gg2024_dhxy@163.com

Abstract. This article focuses on the research of intelligent financial risk identification and early warning models driven by big data platforms. By integrating multiple sources of data and using machine learning algorithms, an intelligent identification and early warning model for financial risks is established. The big data platforms is used to collect, clean, and process various types of financial data, including financial statements, transaction data, and market indicators. Secondly, combined with data mining and model building techniques, a supervised learning based intelligent financial risk identification model is constructed to achieve timely identification of abnormal changes and potential risks. Finally, the effectiveness and practicality of the model are verified through empirical case analysis, providing reliable early warning support for enterprise financial risk management. This study has important theoretical and practical significance in the field of financial risk management, providing new ideas and methods for building an intelligent financial risk warning system.

Keywords: BDP · Intelligence · Financial Risk · Early Warning Models · Data Mining · Machine Learning · Supervised Learning

1 Introduction

With the rapid development and widespread application of big data technology, big data platforms (hereinafter referred to as BDP) are becoming an important driving force for the intelligent and digital transformation of enterprise finance. In today's fast-paced and information explosion economic environment, enterprises are facing increasingly complex and variable financial risks, which pose a huge challenge to the stable operation and development of enterprises. How to effectively utilize BDP technology to achieve intelligent financial risk identification (hereinafter referred to as FRI) and early warning (hereinafter referred to as EW) has become an urgent problem to be solved. Traditional financial risk management relies on manual experience and rulemaking, facing problems such as insufficient data sources, low analysis efficiency, and delayed warning responses. In order to address these challenges, researchers have begun to introduce BDP and machine learning technologies into the field of financial risk management, constructing

intelligent FRI and warning models. BDP not only have the ability to efficiently process massive amounts of data, but also integrate multiple sources of data, explore potential correlations, and provide more comprehensive information support for financial risk analysis; Machine learning algorithms can automatically learn data features and patterns, achieving fast and accurate identification and EW of risks. This article is driven by BDP and explores the research of intelligent FRI and EW models. Combining the characteristics of BDP, different types of financial data, including financial statements, transaction records, market indicators, etc., are integrated to form a complete financial dataset [1]. Through data preprocessing, feature engineering and other technologies, valuable features of the data are extracted, and a supervised learning based FRI model is constructed. This model can automatically identify abnormal changes and potential risk signals, providing timely EW support for enterprise risk management. I hope that this study can provide more intelligent and efficient solutions for enterprise financial risk management, effectively respond to constantly changing risk challenges, enhance the risk identification and EW capabilities of enterprises, and ensure their sustained and stable development.

2 Overview of BDP and Intelligent FRI and EW Models

2.1 Introduction to BDP

As an important infrastructure for implementing big data management and application, BDP are increasingly becoming a support for intelligent decision-making and risk management in enterprises. BDP have high scalability and fault tolerance. Compared with traditional data processing systems, BDP adopt a distributed architecture design, which can flexibly expand computing and storage resources, avoid single points of failure, and ensure the safety and stability of data. The BDP also has diversified data integration capabilities, and enterprises have diverse data sources. The BDP can integrate data from different data sources, achieve data cleaning, conversion, and storage, and provide a unified data view for subsequent data analysis and mining. In addition, the BDP also provides powerful data analysis tools and algorithm support. Through the BDP, enterprises can use advanced algorithms such as data mining and machine learning to discover the patterns and trends behind the data, thereby achieving data-driven intelligent decision-making and risk management. With the changing business environment and rapid growth of data, enterprises need to respond quickly and monitor data changes in real time. BDP can achieve real-time data processing and analysis, ensuring that enterprises have timeliness and flexibility [2]. The application of BDP in the financial field is of great significance. By utilizing BDP, enterprises can more quickly integrate and analyze massive financial data, thereby achieving intelligent FRI and EW.

2.2 Definition and Principle of Intelligent FRI and EW Model

The intelligent FRI and EW model, combined with financial data analysis and prediction methods, can achieve automated identification and timely warning of enterprise financial risks. Specifically, the model can integrate and process financial information from

different data sources, including financial statements, transaction records, and market indicators, through BDP. In the process of integration and processing, the model can also use technologies such as data cleaning, conversion, and storage to ensure data quality and comprehensiveness, and prepare data for subsequent risk identification and EW. The model is based on machine learning algorithms for feature extraction and pattern recognition of financial data, discovering potential correlations and patterns between data. Through supervised learning methods, it can automatically learn normal business rules and abnormal behavior patterns, thereby identifying potential risk signals and abnormal changes. Once data anomalies or thresholds are detected, the model will immediately issue warning signals to remind relevant personnel of the enterprise to take timely response measures to avoid possible financial risks. The intelligent FRI and EW model can also achieve trend analysis and prediction of risk events, providing reference basis for enterprise financial decision-making. By continuously optimizing model parameters and algorithms, the accuracy and practicality of the model can be improved, enabling enterprises to respond more sensitively to risk challenges, ensure financial security and stable development [3].

2.3 The Role of BDP in FRI and EW

BDP provide powerful data processing and analysis tools for the financial sector, promoting intelligent and automated financial risk management. It integrates multiple sources of internal and external data to form a comprehensive view of financial data. During the FRI process, it is possible to collect and summarize massive amounts of data in real-time, avoiding information silos and data fragmentation, and building a solid information foundation for financial risk analysis [4]. The BDP has extremely high efficiency in processing large-scale data, which can quickly clean, convert, and calculate massive financial data, ensuring the real-time and accuracy of FRI and EW models, and helping enterprises quickly respond to risks and reduce losses. It supports data mining and machine learning algorithm applications, can mine data patterns, identify potential financial risks, and achieve risk prediction and early warning through algorithm learning of historical data. In addition, BDP provide flexible data visualization and reporting tools to visually present complex financial data, helping enterprise managers understand financial and risk situations, make quick decisions, manage risks and strategic planning, improve risk identification and early warning efficiency, and ensure the financial security and stable development of enterprises. The workflow of the financial analysis platform based on data mining technology is shown in Fig. 1.

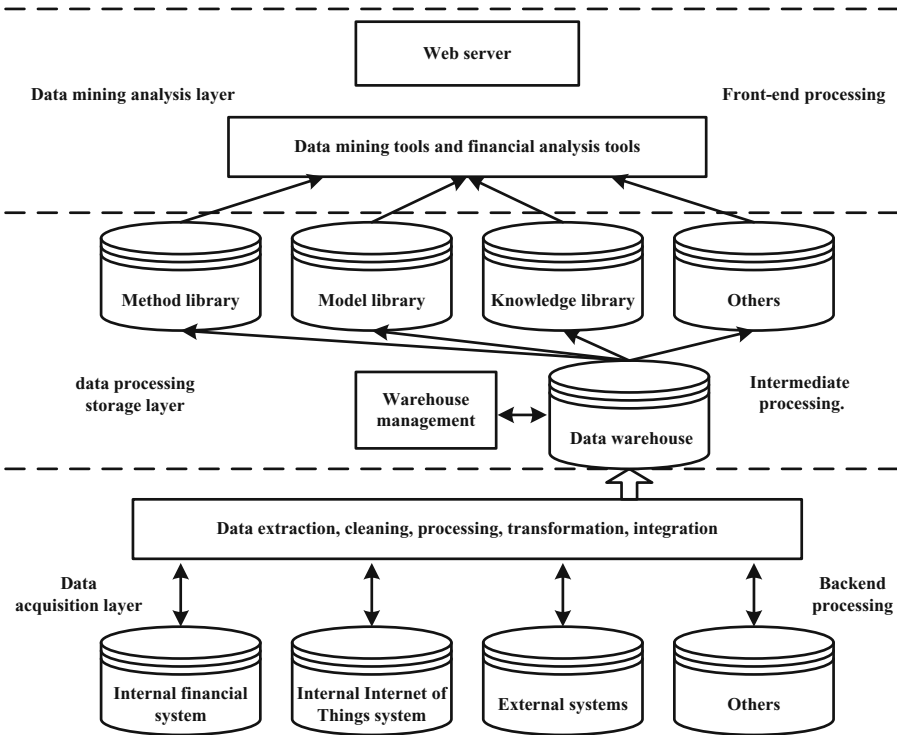


Fig. 1. Workflow of a financial analysis platform based on data mining techniques.

3 Construction of an Intelligent FRI and EW Model

3.1 Data Collection and Preprocessing

Building intelligent FRI and EW models, data collection and preprocessing are key steps. It is necessary to collect data from multiple sources such as internal financial systems, market and industry data, and use API interfaces and data warehouses to ensure timely and complete data, while strictly adhering to data security and privacy regulations. When cleaning data, handle missing values, outliers, and duplicate values, and improve data quality and reliability through operations such as deduplication and imputation. Data conversion and integration, format conversion, standardization, etc., integrate multi-source data to form a unified view, and provide comprehensive information for the model. Feature engineering uses statistical and data mining methods to screen and construct financial risk related features, enhancing the predictive ability of the model [5]. Divide the training set and testing set for model training and parameter optimization to verify predictive performance. These steps lay a solid data foundation for building intelligent FRI and EW models, improving model accuracy and practicality, achieving intelligent risk management and decision support. The data collection process is shown in Fig. 2.

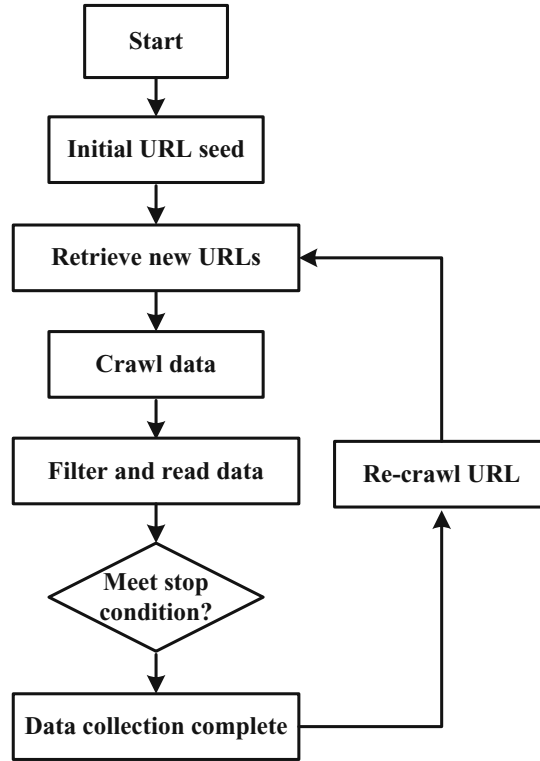


Fig. 2. Data collection process.

3.2 Feature Extraction and Selection

By extracting and selecting appropriate features, the accuracy and generalization ability of intelligent FRI and EW models can be improved. In the feature extraction stage, representative and relevant features need to be extracted from the original financial data, such as financial indicators, account transaction information, market indicators, etc. Through statistical analysis, data mining and other technologies, features closely related to financial risks can be obtained. Feature construction should combine and derive the extracted features based on actual needs and data characteristics to construct new features, such as calculating ratios, growth rates, etc., to enhance the sensitivity of the model to financial risks. Feature selection requires screening and selecting features that have a significant impact on the model's prediction of financial risk, removing redundant or irrelevant features. Common feature selection methods include filtering, wrapping, and embedding. In the process of feature selection, high-dimensional data may face challenges. To reduce computational complexity and avoid overfitting, feature dimensionality reduction techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) are used to compress and simplify the data. In the process of establishing an intelligent FRI and EW model, it is necessary to continuously evaluate and adjust the effectiveness of features to ensure that the selected features can

provide sufficient information and effectively support model training and prediction [6]. Through cross validation, model evaluation and other methods, the quality of features and their impact on model performance are evaluated, and the strategy for feature selection is continuously optimized to provide effective feature support for the intelligent FRI and EW model, improving the accuracy and robustness of the model. In a big data cloud computing environment, the financial statement data can be extracted using in-memory encoding, and the fused feature solution vector can be obtained through the fusion of financial statement data, computed as shown in Eq. (1).

$$R = \text{sgn} \left(\sum_{i=1}^n t_i P_i x_i + \delta \right) \quad (1)$$

In the equation, P_i represents the feature quantity matched for the financial statement data.

3.3 Model Training and Optimization

By training and optimizing models, the predictive accuracy and generalization ability of intelligent FRI and EW models can be improved, achieving accurate identification and effective EW of financial risks. During the model training phase, it is necessary to select algorithms suitable for FRI and EW, such as decision trees, support vector machines, neural networks, etc. Different algorithms have different characteristics and applicable scenarios, and the most suitable algorithm needs to be selected for modeling based on data characteristics and actual needs. By using annotated training datasets, the selected algorithm needs to be trained to learn the patterns and patterns of the data [7]. Model training requires adjusting the parameters of the algorithm and setting appropriate model structures to achieve the best performance of the model. After the model training is completed, it is necessary to evaluate the model using test data to evaluate its accuracy, precision, recall, and other performance indicators. Based on these evaluation results, the model should be optimized, including parameter tuning, feature selection, and model structure adjustment. Multiple high-quality models should be considered for integration to improve overall prediction accuracy and robustness. A highly accurate and stable intelligent FRI and EW model should be constructed to help enterprises identify and prevent potential financial risks in a timely manner, ensuring financial security and sustainable development. The framework for mining and analyzing financial statement data is shown in Fig. 3.

3.4 Evaluation and Validation of EW Models

By comprehensively evaluating and validating the model, the effectiveness and reliability of the intelligent FRI and EW model can be verified, ensuring its accuracy and stability in practical applications. In the model evaluation stage, various evaluation indicators need to be used to comprehensively evaluate the model, such as accuracy, recall, precision, F1 score, etc., to help evaluate the recognition and EW effect of the model on financial risks, evaluate the comprehensiveness and accuracy of the model. In the model

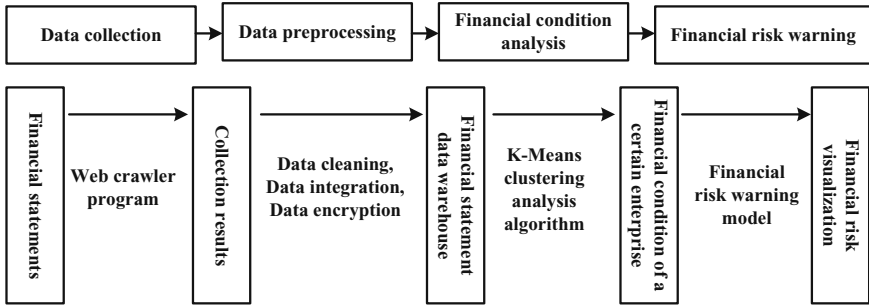


Fig. 3. Framework of financial statement data mining and analysis model.

validation process, independent test datasets need to be used to validate the model, test the predictive performance of the model on unknown data, and verify the reliability and practicality of the model [8]. Error analysis is required in model evaluation and validation. By comparing the predicted results of the model with the actual situation, the reasons and types of errors are analyzed. Through in-depth analysis of the errors, problems and areas for improvement in the model are identified, further improving the accuracy and effectiveness of the model's predictions. In the process of model evaluation and validation, the performance of different models can be compared, the optimal model can be selected for application, and the most suitable model for specific financial risk scenarios can be found to improve the accuracy and applicability of EW models. To verify the effectiveness of the EW model in practical business, some financial risk cases can be selected for practical verification of the model. Through the prediction and application of actual cases, the prediction effect of the model in real scenarios can be tested, further verifying the practicality and effectiveness of the model.

4 Application Cases of Intelligent FRI and Warning Models Driven by Four BDP

4.1 Case Selection and Data Preparation

Selecting appropriate cases and preparing data are crucial for the effectiveness and practicality of intelligent FRI and EW models. When selecting cases, focus on the characteristics and typicality of financial risks, select cases with rich financial data, clear risk factors, and representativeness to reflect different types and scenarios of financial risks, and enhance the universality and adaptability of the model [9]. In the data preparation stage, a large amount of financial data is collected from multiple channels such as internal systems, financial statements, and industry databases, covering various aspects such as financial indicators, transaction records, and market data. Next, perform data preprocessing to handle missing values, remove outliers, and improve data quality and availability. To address the issue of inconsistent data formats and structures across different data sources, carry out data integration and conversion. It is also necessary to extract features closely related to financial risks and divide the dataset, usually in a ratio of 70% training set and 30% testing set, to ensure that the model is trained and validated on sufficient and accurate data, and to improve the accuracy and reliability of the model.

4.2 Actual Application Process and Effect

The process and effectiveness of practical application are crucial in the application of intelligent FRI and EW models driven by BDP [10]. It is necessary to deploy the developed intelligent FRI and EW model to the BDP and conduct testing to verify the effectiveness and practicality of the model. The model should be applied to the financial data of enterprises to identify and EW financial risks. By comparing the warning results of the model with actual risk events, the accuracy and predictive ability of the model should be evaluated. Based on the feedback and results of actual application, the model should be adjusted and optimized to further improve its effectiveness and performance. In the process of model application, it is necessary to continuously monitor the application effect and performance of the model, timely detect and solve the problem of inaccurate or missed model warnings, in order to ensure the effectiveness and reliability of the model in practical applications. By summarizing and analyzing the effectiveness and performance of the model in practical applications, the actual application effectiveness and value of the model can be evaluated, lessons learned can be summarized, and references can be provided for further improvement and optimization of the model. It can be effectively applied in practical business scenarios, providing accurate and reliable FRI and EW capabilities, providing important support for enterprise risk management and decision-making, and helping enterprises achieve risk prevention and stable business development.

4.3 Outlook for Future Applications

With the continuous development of big data technology and artificial intelligence, intelligent FRI and EW models will have a broader application prospect driven by BDP. The models will continue to expand to a wider range of fields. In addition to being applied in traditional enterprise finance, they can also be extended to various financial fields such as finance, insurance, and investment, achieving more comprehensive financial risk coverage. The models can also be combined with intelligent systems in other fields to achieve comprehensive risk management and improve intelligence [11]. The future intelligent FRI and EW model will gradually achieve intelligent decision-making and self-learning capabilities through cutting-edge technologies such as deep learning and natural language processing. It can dynamically adjust the EW model based on real-time data and continuously learn and optimize it in practice to improve the accuracy and timeliness of identification and EW. It will better integrate with practical applications. At the same time, the model will be more closely aligned with the actual needs and business scenarios of enterprises, and closely integrated with BDP to achieve personalized customization and application, providing accurate decision support for enterprises. BDP provide stronger data support and computing capabilities, more data interfaces and tools, and further enhance the practicality and effectiveness of the model. In the future, the intelligent FRI and EW model will continue to iterate and optimize, strengthen the combination with practical experience, improve the performance and intelligence level of the model, meet the needs of enterprises for financial risk management. Driven by BDP, it will usher in a wider range of application areas, improve the level of intelligence, and be closer to practical applications. It will be better integrated with BDP, continuously optimized and improved, providing more comprehensive and effective solutions

for enterprise financial risk management, bringing more opportunities and challenges to risk management in the financial field, and promoting the widespread application and development of intelligent technology in the financial field [12]. The overall conceptual process flow is illustrated in Fig. 4.

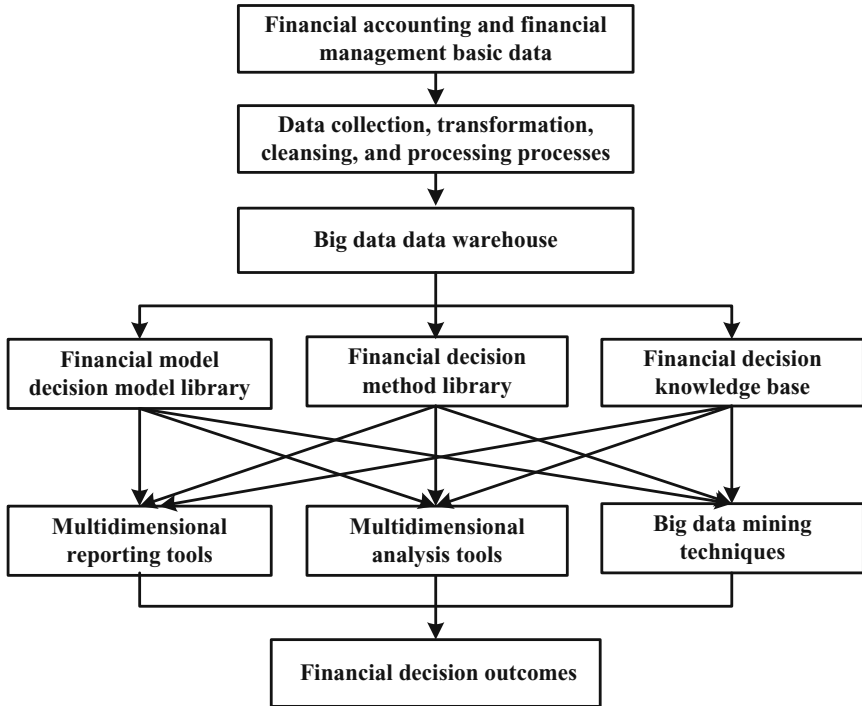


Fig. 4. The overall architecture of an intelligent financial decision support system.

5 Conclusion

This study constructed an intelligent FRI and EW model based on a BDP, achieving automated processing and analysis of financial data. Through the prediction and identification of the model, potential financial risks can be quickly identified and warning signals can be issued in a timely manner, helping enterprises avoid or reduce losses caused by risks and improving financial security. This study provides effective financial risk management tools and decision support for enterprises. Through intelligent FRI and EW models, enterprises can better monitor and manage financial risks, adjust risk prevention strategies in a timely manner, and improve the efficiency and timeliness of risk handling. This will help companies make wiser financial decisions, ensuring long-term development and stable operation. The model has successfully identified multiple types of risks in practical applications and provided timely and effective warnings, indicating

the potential and value of the model in intelligent FRI and warning driven by BDP. The warning accuracy of the model needs to be improved in certain specific situations, and further optimization of the model algorithm and parameters is needed to improve the performance and effectiveness of the model. This study provides useful empirical analysis and practical application for the research of intelligent FRI and EW models under BDP, and presents good application prospects and promotion value. Based on continuous deepening of research and practice, the intelligent FRI and EW model will provide stronger support for enterprise risk management and decision-making, promote sustainable development and financial stability of enterprises.

References

1. Yang, X.: Intelligent integration of enterprise financial big data based on association rule mining. *China Manag. Informatiz.* **26**(21), 61–64 (2023)
2. Zhou, H.: Analysis of financial management strategies for companies based on data mining techniques. *Bus. Obs.* **9**(27), 101–104 (2023)
3. Zhou, W.: Financial statement data mining and analysis in the era of big data. *China New Technol. New Prod.* **5**(15), 139–141 (2023)
4. Liu, K.: Cloud-based financial data mining methods for enterprises. *Inf. Comput. (Theor. Ed.)* **35**(14), 203–205 (2023)
5. He, X.: Overview of financial fraud detection based on data mining techniques. *Digit. Technol. Appl.* **41**(06), 4–6 (2023)
6. Gao, N.: Data mining and analysis of enterprise financials in the context of big data applications. *China New Technol. New Prod.* **31**(12), 133–136 (2023)
7. Wang, Q., Guan, H.: Financial fraud data mining in listed companies: analysis based on time dimension. *Adm. Assets Financ.* **23**(3), 96–98 (2023)
8. Dou, L.: Enterprise financial accounting information management system based on big data mining. *China Manag. Informatiz.* **25**(22), 74–76 (2022)
9. Zhang, Y., Yuan, J.: Application and challenges of big data mining in corporate financial analysis. *Mod. Bus. Mark.* **8**(17), 117–119 (2022)
10. Feng, Y., Guo, X.: Research on financial data mining and application based on big data thinking. *Sci. Technol. Econ. Mark.* **11**(7), 22–24 (2022)
11. Wang, H.: Application research of data mining and analysis in intelligent finance. *Financ. Account. Stud.* **15**(24), 53–54 (2021)
12. Ye, F.: Application of data mining in intelligent financial decision support systems. *J. Langfang Norm. Univ. (Nat. Sci. Ed.)* **15**(02), 38–40 (2015)



Analyzing the Impact of Digital Transformation on Earnings Management Based on Stata Software Taking H Company as an Example

Zhihua Zhang and Jiaqi Ma(✉)

School of Economics and Management, Yanbian University, YanjiJilin 133000, China
614430812@qq.com

Abstract. Earnings management behavior in enterprises is a concern in the academic community. In actual operations, due to factors such as information asymmetry, management may use earnings manipulation to distort the true financial situation of the enterprise and embellish financial reports in order to alleviate financing pressure and achieve performance commitments. This will mislead the public's understanding of corporate performance, seriously harm the interests of accounting information users, and reduce the effectiveness of the capital market. Therefore, it is crucial to explore how to effectively suppress earnings management behavior. With the tight integration of enterprise operations and digital technology, the internal governance effectiveness of digital technology can be demonstrated, thereby controlling the selectivity of enterprise accounting policies and influencing earnings management behavior. This article takes H Company as an example, based on its annual report data from 2018 to 2022, to explore its earnings management issues from the perspectives of accrual-based and real earnings management, and proposes countermeasures and suggestions for digital transformation.

Keywords: Digital Transformation · Big Data · Accounting Information Quality · Suppress · Accrual-based Earnings Management · Real Earnings Management

1 Introduction

In the past, the academic community mainly analyzed the issue of corporate earnings management based on internal control, governance structure, tax planning, etc., and there was little research on the impact mechanism of designing big data applications. This article discusses from the perspective of data empowerment, using Stata software to analyze how data can be used to control earnings management for public companies.

This article takes H company as the starting point, constructs indicators of big data technology application involved in H Company's financial data from 2018 to 2022, explores the impact of information sharing and big data technology application on H Company's earnings management and its regulatory effect on its dynamic ability, and tests the impact mechanism of digital transformation on H company's accrual-based and

real earnings management. From the two aspects of internal governance and external governance, seek the conditions that are conducive to the role of modern technology, and provide reference for other enterprises.

Compared with existing research, this article will comprehensively examine the role of digital operations in corporate earnings management from the perspectives of accrual-based and real earnings management. With the rapid development of digital economy, science and technology are one of the important factors affecting productivity, and modern technology has gradually become the foundation and core of all digital economy technologies. In addition, financial data is the most important information of company. Therefore, this article intends to conduct research from both theoretical and empirical perspectives.

2 Introduction to H Company's Basic Information

2.1 Overview of H Company

H Company's main business includes the production and sales of forestry, door products, mineral water, and seedlings, as well as landscape engineering construction and landscape design. As a listed company, H Company has an internal shareholders' meeting, board of directors, and supervisory board.

Since 2020, the company has gradually undergone digital transformation through software and hardware upgrades, and improved digital industry construction, which has improved the company's earnings management status.

2.2 Financial Status of the Company

Earning Power

This section mainly selects three main profitability indicators to analyze H company.

The operating profit margin of H Company sharply decreased to -16.97% in 2015. This is related to the company's main business scope and nature of operation: In 2015, the government issued the "State owned Forest Farm Reform Plan", which proposed to promote the transformation of the forestry development model from wood production as the main focus to ecological restoration and construction as the main focus, from using forests to obtain economic benefits to protecting forests and providing ecological services as the main focus, and to protect and develop forest resources. A large part of China's forest resources come from state-owned forest farms [1]. The introduction of this policy means that the country has increased the protection of forest resources, and the development of forest industry enterprises depends on forest resources. Therefore, the cost of forest industry enterprises has further increased, and the profit space has been further compressed, resulting in a decrease in the company's profitability. But the advent of challenges often comes with opportunities, and the introduction of this policy also brings more opportunities and possibilities for the development of the forestry tertiary industry.

H Company's operating profit margin has slowed down from 2015 to 2018. This is closely related to the risk response strategy and strategic adjustment of the enterprise.

Until 2019, H Company's profitability indicators experienced significant fluctuations and a significant decline. In 2019, the operating profit margin, sales net profit margin, and total asset net profit margin decreased significantly. This shows that the profitability of H company suddenly deteriorated [2].

Since 2020, the company's profitability indicators have shown a significant improvement. Under the guidance of China's forestry policies, the company continuously optimizes its industrial structure, actively seeks new breakthroughs and innovations, accelerates the development speed of the forestry tertiary industry, and gradually slows down the development speed of the forestry primary and secondary industries.

Operating Capacity

The data show that the accounts receivable turnover rate of H company in 2017 has shown a significant improvement compared to 2015, but the current asset turnover rate and total asset turnover rate have shown a downward trend. This indicates that there may be some non-performing assets in the company, and the operating capacity of these non-performing assets is gradually declining, offsetting or even exceeding the positive operating capacity effects generated by high-quality internal assets.

By analyzing the company's financial statements, it can be found that the company recognized large accounts receivable in 2019 and 2020. The book value of accounts receivable increased by 143.21% year-on-year in 2019 and 72.01% year-on-year in 2020. It is precisely because of the recognition of large accounts receivable that the turnover rate of accounts receivable and current assets has significantly decreased in the past two years. The reason for the increase in total asset turnover may be due to the company's recognition of significant asset losses in 2019, resulting in a significant decrease in the book value of long-term equity investments, thereby affecting the total assets.

From 2020 until 2022, the company's three profitability indicators have all been declining [3]. From the relevant indicators, there is still room for improvement in the company's operational capacity, which requires close attention from enterprise management personnel.

Debt Repayment Capabilities

In terms of short-term solvency, this section selects three relevant financial indicators for analysis. From 2013 to 2019, H Company's current ratio and quick ratio showed an overall downward trend, and its short-term solvency gradually decreased. Between 2020 and 2022, the company's short-term solvency index has increased, but this does not necessarily mean an improvement in the company's short-term solvency. Overall, there has been no significant improvement in H Company's short-term solvency.

In terms of long-term solvency, this section selects asset liability ratio and equity ratio as analysis indicators. Between 2013 and 2017, the asset liability ratio and equity ratio indicators of H Company decreased significantly compared to before. This means that after completing business transformation, the company's long-term debt paying ability has improved, and the debt paying pressure has gradually decreased.

Compared to before, the asset liability ratio and equity ratio have significantly increased in 2019, indicating that the company is facing enormous debt repayment pressure. By analyzing the company's balance sheet, we can find that from 2017 until 2020,

the company began issuing bonds and the amount of share capital gradually increased. At the same time, the company also confirmed significant investment losses, resulting in a significant decrease in the book value of its long-term equity investments, thereby affecting the company's total assets [4]. And the huge investment losses also affected the net profit and undistributed profit of the year, significantly reducing owner's equity, resulting in an increase in the proportion of liabilities and increasing the company's long-term debt repayment pressure.

From 2020 to 2022, the company's asset liability ratio and equity ratio have significantly improved compared to 2019, indicating that the company's debt paying ability is gradually recovering and its main business has not been significantly impacted.

3 Implementation of Earnings Management Model Based on Stata Software

3.1 Research Design

This article selects relevant financial data of H Company from 2018 to 2022 for calculation. At the same time, the modified Jones model commonly used in domestic empirical research is used to measure the degree of accrual-based earnings management, and the Roychowdhury (2006) model is used to measure the true degree of earnings management. By comparing the cash flows, operating income and costs, inventory levels, and period expenses of adjacent accounting periods, abnormal cash flow from operating activities, abnormal production costs, and abnormal expenses are calculated, and the above three measurement indicators are summed to measure the true degree of earnings management. We identify whether the company has adjusted its actual production and operating activities through sales manipulation, production manipulation, and expense manipulation. Finally, we use Stata14.0 software to evaluate whether the company has positive earnings management behavior [5].

3.2 Research Process

Accrual-Based Earnings Management

The accrual profit separation method is divided into discretionary accruals and non discretionary accruals. Among them, the average change in "non discretionary accruals" is almost zero, so the change in average cumulative profit almost reflects the average change in "discretionary accruals". Jones model evaluates the profit margin of "non-discretionary accruals" to determine the profit margin of "non-discretionary accruals". In future exploration, optimization was carried out on the basis of the Jones model to construct a more accurate model.

The revised Jones model suggests that the increase in operating revenue should be deducted from the increase in accounts receivable when calculating non-discretionary accruals [6].

$$\begin{aligned} & \text{Non manipulative accrued surplus} \\ &= \beta_0 \frac{1}{\text{Total assets at the end of the previous year}} \end{aligned}$$

$$\begin{aligned}
& + \beta_1 \frac{\text{Changes in operating income} - \text{Changes in accounts receivable}}{\text{Total assets at the end of the previous year}} \\
& + \beta_2 \frac{\text{Current fixed assets}}{\text{Total assets at the end of the previous year}}
\end{aligned} \quad (1)$$

Considering the cross-sectional correction of performance in the Jones model: ROA is placed as a control variable in the Jones model to analyze the impact of performance on discretionary accruals. Therefore, a constant term is introduced in the model, where ROA can also be used for the t-period.

$$\begin{aligned}
& \frac{\text{Accrued surplus}}{\text{Assets at the end of the previous year}} \\
& = \beta_0 \beta_1 \frac{1}{\text{Total assets at the end of the previous year}} \\
& + \beta_2 \frac{\text{Changes in operating income}}{\text{Total assets at the end of the previous year}} \\
& + \beta_3 \frac{\text{Current fixed assets}}{\text{Total assets at the end of the previous year}} + \beta_4 \text{ROA}_{i,t-1} \\
& + \text{Abnormal changes}
\end{aligned} \quad (2)$$

In the revised Jones model, ROA is added to calculate the non-discretionary accruals, as shown in formula (3). The discretionary accruals is the total accrued profit minus the non-discretionary accruals, as shown in formula (4).

$$\begin{aligned}
& \text{Non manipulative accrued surplus} \\
& = \beta_0 + \beta_1 \frac{1}{\text{Total assets at the end of the previous year}} \\
& + \beta_2 \frac{\text{Changes in operating income} - \text{Changes in accounts receivable}}{\text{Total assets at the end of the previous year}} \\
& + \beta_3 \frac{\text{Current fixed assets}}{\text{Total assets at the end of the previous year}} + \beta_4 \text{ROA}_{i,t-1}
\end{aligned} \quad (3)$$

$$\begin{aligned}
& \text{Manipulative accrued surplus} \\
& = \frac{\text{Accrued surplus}}{\text{Total assets at the end of the previous year}} \\
& - \text{Non manipulative accrued surplus}
\end{aligned} \quad (4)$$

In the model, when a listed company conducts positive accrual earnings management, the manipulative accrual earnings are greater than 0, and vice versa, it is less than 0, indicating reverse accrual earnings management. When studying the degree of earnings management, absolute values are taken for the results of manipulative discretionary accruals [7].

Real Earnings Management

To avoid losses, the company will increase profits from actual business activities, such as increasing sales in a short period of time through discounts. This article uses a model to conduct regression analysis from three aspects: abnormal operating cash flow, abnormal production costs, and discretionary expenses, in order to estimate the true revenue

management of the enterprise.

$$\begin{aligned} \text{Comprehensive indicators} = & (-1) * \text{Discretionary control expenses} + (-1) * \text{Net} \\ & \text{cash flow from manipulative operations} + \text{manipulative production costs} \end{aligned} \tag{5}$$

Manipulative production cost is a positive indicator, while manipulative operating net flow and manipulative discretionary expenses are negative indicators. The larger the comprehensive indicator value, the more actively a company engages in positive earnings management activities. When a company uses excessive production to increase profits, it also leads to lower cash flows. At this time, the cost of abnormal products overlaps with abnormal cash flows [8]. Therefore, some literature does not add up all three indicators when calculating the total true earnings management level.

4 Analysis of Model Implementation Results

4.1 Analysis of Accrual-Based Earnings Management Results

When the discretionary accruals is greater than 0, the larger the discretionary accruals, the greater the degree of positive earnings management; When the discretionary accruals is less than 0, the smaller the discretionary accruals, the greater the degree of negative earnings management. The following table shows the results of calculating the corresponding earnings management through the model.

Table 1. H Company’s manipulative discretionary accruals from 2020 to 2022

Time	Discretionary accruals	Industrial average of positive discretionary accruals	Industrial average of discretionary accruals
2022	0.0298	0.0698	0.0075
2021	0.0601	0.0496	0.0186
2020	0.0556	0.0599	0.0089

According to Table 1, it can be seen that the discretionary accruals of H Company from 2020 to 2021 is positive and higher than the industry average, which is equivalent to the industry’s positive earnings management manipulative accrued surplus average. It can be inferred that H Company may have positive earnings management behavior.

4.2 Analysis of Real Earnings Management Results

Abnormal Cash Flows from Operating Activities

When the abnormal cash flow from operating activities is negative, the smaller the value, the more serious the positive real earnings management behavior; When the abnormal cash flow from operating activities is positive, the larger the value, the more

Table 2. H Company's abnormal CFO from 2020 to 2022

Time	Abnormal CFO	Industrial average of positive abnormal CFO	Industrial average of abnormal CFO
2022	−0.0209	−0.0710	−0.0004
2021	−0.1102	−0.0489	−0.0005
2020	−0.0404	−0.0509	−0.0003

severe the negative real earnings management behavior. The following table shows the results of calculating abnormal cash flows from operating activities using a model.

According to Table 2, the abnormal cash flow from operating activities of H Company from 2020 to 2021 was negative, and this value was lower than the industry average, which is equivalent to the average abnormal cash flow from operating activities under positive earnings management in the industry. It can be inferred that H Company may have positive earnings management behavior.

Abnormal Production Costs

When the abnormal production cost is positive, the higher the value, the more severe the degree of positive earnings management; When the abnormal production cost is negative, the smaller the value, the more severe the negative earnings management. The following table shows the results of calculating abnormal production costs through a model.

Table 3. Abnormal production costs of H Company from 2020 to 2022

Time	Abnormal production costs	Industrial average of positive abnormal production costs	Industrial average of abnormal production costs
2022	−0.0356	0.0906	0.0003
2021	0.1578	0.0621	0.0005
2020	0.0612	0.0600	0.0002

According to the data in Table 3, except for 2022, the abnormal production costs of H Company in all other years are positive, and this value is higher than the industry average, which is equivalent to the average abnormal production costs of positive earnings management in the industry. It can be inferred that H Company may have positive earnings management behavior.

Abnormal Discretionary Expenses

When the abnormal discretionary expenses are negative, the smaller their value, the more severe the positive earnings management behavior; When the abnormal discretionary

expenses are positive, the larger their value, the more severe the negative earnings management behavior. Thus, identifying abnormal discretionary expenses can measure and evaluate the degree of earnings management. The following is the result of calculating abnormal costs through the model [9].

Table 4. H Company’s abnormal discretionary expenses from 2020 to 2022

Time	Abnormal discretionary expenses	Industrial average of positive abnormal discretionary expenses	Industrial average of abnormal discretionary expenses
2022	−0.0141	−0.0401	−0.0003
2021	−0.0387	−0.0346	−0.0004
2020	−0.0241	−0.0306	−0.0004

According to the data in the Table 4, H Company’s abnormal production cost values from 2020 to 2021 were lower than the industry average, which is equivalent to the industry average of abnormal discretionary expenses for positive earnings management. This indicates that H company may have engaged in positive earnings management behavior.

In statistical analysis, data from the past five years were used and robustness tests were conducted. Due to space limitations, only data from the past three years have been selected for display here.

5 Research Conclusions and Policy Recommendations

5.1 Research Conclusion

This article selects H Company and conducts research on its earnings management behavior. Build a model based on existing theories and use Stata software to calculate industry data and relevant data of H Company. The final conclusion is that H Company may have positive earnings management behavior. Through the previous analysis, it can be found that although the company has earnings management issues, with the gradual improvement of its digital transformation, the degree of earnings management is decreasing year by year. Therefore, this indicates that digital transformation may have a positive impact on the company’s earnings management issues.

5.2 Improvement Suggestions

Improve the Company’s Digital Governance Structure

Enterprises should attach great importance to the improvement of their corporate governance structure. In the process of financial information supervision, it is necessary to achieve “problem to person”, clarify the responsibilities of each department and employee, establish sound rules and regulations, and reduce the operational space and

possibility of earnings management [10]. A strict corporate governance model and a sound reward and punishment system can not only effectively regulate earnings management behavior, but also further cultivate employees' sense of belonging and responsibility towards the company. The company can simultaneously build a digital audit platform and internal audit institutions, establishing the first line of defense within the company.

Using Digital Means to Improve the Company's Internal Control

Good internal control quality can ensure the reliability and authenticity of financial information disclosure. The improvement of the internal control system can further ensure the strict implementation of the production and operation links. The deepening of digital transformation has greatly improved the efficiency of enterprise information processing, accelerated the flow of information within the organization, established a digital monitoring system by using digital transformation technology, formed a digital internal governance environment, and enhanced the internal control ability of enterprises. In the digital monitoring system, managers and employees can clearly grasp the flow of company funds, clear financing, investment and operating activities of the proportion of funds. Secondly, real-time data enables managers to keep abreast of sales and inventory, adjust sales strategies in a timely manner, improve the decision-making process based on capital structure and risk analysis, reduce daily business and investment risks, and strengthen the orderliness and scientificity of organizational business behavior.

Improve Information Transparency

Strengthen cooperation with external authoritative forces, constantly adjust measures, and explore the optimal solution for transformation. Relying on digital technology, corporate financial accounting information can be transmitted to external stakeholders through shared equipment and services, which promotes the improvement of external supervision, forces managers to converge violations, and reduces the degree of earnings management. Promote the management of the enterprise to respond well to the market and the long-term development of the enterprise, and reduce the willingness to make earnings management activities that damage the value of the enterprise.

Improve Employee Digital Literacy

In the process of digital transformation, enterprises should realize the importance of 'data' elements and focus on improving the digital skills of employees, especially executives. With the help of data to convey a good operating signal to the outside world, make full use of the obtained data information to meet customer needs. The application of digital technology enables more and more links in enterprise operation to be transformed into data, improves the sharing degree of internal information, and enhances the feasibility and diversity of information disclosure. In terms of employee skills, enterprises can comprehensively use offline training and online learning to improve employees' digital technology application ability, and can also recruit high-quality talents in big data analysis, computer and other related majors to obtain the ability required for digital transformation of enterprises. In terms of data analysis, enterprises should pay attention to the collection of customer information, understand and deal with customer needs, carry out R & D innovation around the analysis results of customer information, and

obtain more customer recognition. Only by giving full play to the key role of data, can enterprises alleviate problems such as information asymmetry through digital transformation, improve the quality of enterprise information, and then improve the competitive vitality in the capital market.

Acknowledgements. Fund Project: Yanbian University 2020 Educational Science Research Project (ZD2020011).

References

1. Aziza, N., Ahmed, N.S., Khatib, S.F.A., et al.: Illuminating the shadows: a systematic review of earnings management practices in family-owned enterprises and future research directions. *J. Bus. Socio-Econ. Dev.* **4**(4), 340–358 (2024)
2. Jin, X., Mirza, S.S., Huang, C., et al.: Digital transformation and governance heterogeneity as determinants of CSR disclosure: insights from Chinese A-share companies. *Corp. Gov.: Int. J. Bus. Soc.* **24**(6), 1314–1336 (2024)
3. Hlel, K.: Tax avoidance and earnings management: evidence from French firms. *Int. J. Financ. Eng.* (2024). (Prepublish)
4. Lian, L., Zhao, D., Dong, Q.: Fund holdings and real earnings management: an empirical analysis of the Chinese A-share market. *Heliyon* **10**(16), e36194–e36194 (2024)
5. Mouttaqui, E.H., Gleason, K., Jiraporn, P., et al.: How does corporate diversification affect earnings management? A path analysis approach. *Int. Rev. Econ. Financ.* **95**, 103455 (2024)
6. Aljifri, K., Elrazaz, T.: Effect of earnings management on earnings quality and sustainability: evidence from gulf cooperation council distressed and non-distressed companies. *J. Risk Financ. Manag.* **17**(8), 348 (2024)
7. Wang, L., Hou, S.: The impact of digital transformation and earnings management on ESG performance: evidence from Chinese listed enterprises. *Sci. Rep.* **14**(1), 783 (2024)
8. Liao, K., Liu, H., Liu, F.: Digital transformation and enterprise inefficient investment: under the view of financing constraints and earnings management. *J. Digit. Econ.* **2**, 289–302 (2023)
9. Fang, Q., Yu, N., Xu, H.: Governance effects of digital transformation: from the perspective of accounting quality. *China J. Account. Stud.* **11**(1), 77–107 (2023)
10. Jordan, E.C., Clark, J.S., Anderson, M.: Unusual digital patterns in EPS: evidence on the association between earnings management and company characteristic (2011)



Decision Support System for Engineering Project Management Based on Data Mining

Dayong Zhang and Rongwen Chen^(✉)

Changjiang Survey Planning Design and Research Limited Company, Wuhan, China
449071973@qq.com

Abstract. With the increasing complexity of construction projects, enhancing decision-making and managing operational risks have become significant challenges for enterprises. This paper designs and implements an engineering decision support system based on data analysis and machine learning, aimed at providing scientific decision analysis support to enterprises. The system integrates knowledge from engineering management with advanced data mining techniques to construct a series of decision models for cost prediction, schedule control, and quality analysis. These models are efficiently deployed using microservices architecture and offer interactive visual decision interfaces. Comprehensive testing demonstrates the system's excellent functionality, high concurrency processing capabilities, and scalability. In practical applications, the system significantly improves decision quality, reduces operational costs, promotes the transformation of decision-making from empirical to data-driven and model-driven approaches, thereby contributing robust technical support to the high-quality development of the construction industry.

Keywords: Data Mining · Cost Control · Schedule Management

1 Introduction

Decision management in construction projects has, for a long period, been a challenge for companies in the constructed environment. Conventional decision processes, in most instances, relied on expert experiential judgment and could not effectively comprehend the many factors of uncertainty involved in complex construction projects. Conventional decision processes, in most instances, translated into a plethora of complications such as, but not limited to, cost overruns, schedule slips, and many defects in terms of quality. Realizing such failures, in this work of research, a decision support system for engineering is designed and implemented. Development of the system was necessitated by the imperative to present construction companies with model-based and data-driven scientific tools for decision-making in a view to overcoming such complications. With its development resting on fusing expert domain expertise in construction management with cutting-edge artificial intelligence technology, the system is ideally placed to extract rule patterns from a variety of sources of engineering information [1]. Extraction involves

developing a sequence of decision support models specifically designed for key dimensions of construction project management such as cost management, schedule management, and assurance of quality. Efficient use of such models is then in deploying them in microservice form to present decision-makers with analytic reports and prediction simulation output through interactive visualization interfaces. With such a development, through its design, the system is placed to inform scientific decision processes with real-time information and prediction capabilities. With such an improvement, decision accuracy is immensely enhanced, and with it, comes a much-desired revolution in decision-making paradigms in companies. It transforms conventional empirical approaches to decision-making to model and data-driven approaches, and ones that are more robust and fact-based. The system implementation thus provides construction enterprises with a powerful technical supporting system, one that is essential for high-quality development in the construction industry. It is a significant development in the industry, one that leverages cutting-edge technology to provide solutions to age-old construction project management problems.

2 Construction of Data Mining Models

2.1 Algorithm Selection

In the engineering project management decision support system, a significant portion of which consists of the module for managing and analyzing data, careful selection of a variety of algorithms for data mining have been included in it. For selection, these algorithms have been chosen for their versatility in a range of scenarios for a task and for types of processing present in big, complex datasets. For classification, a routine activity in decision making in engineering project management, specifically, decision trees and random forests have been ranked high in selection. Several factors underlay our preference for them. First, these algorithms have high efficiency, capable of processing and classifying big volumes of information in a timely and efficient manner, in terms of computational resources. Second, interpretability is a significant advantage, for easier comprehension and interpretability of justification for a classification model's prediction, even for a domain expert, in critical decision scenarios. Third, these algorithms have high robustness in case of outliers, a prevalent feature in real datasets. That feature lessens distortions and inaccuracies, such distortions and inaccuracies could introduce, in case such outliers go unprocessed appropriately. Thus, with these algorithms included, an improvement in the system's capability to produce correct, interpretative, and robust decision support in engineering project management is hoped for. For instance, in predicting the risk of construction schedule delays based on historical data, we constructed a random forest classification model achieving an accuracy of 83.7%. Meanwhile, for regression tasks, algorithms such as Support Vector Regression (SVR) and Gradient Boosting Decision Trees (GBDT) are employed to predict continuous variables like project costs and durations [2]. Table 1 presents training and validation scores of a GBDT model for cost prediction.

For association rule mining tasks, we utilized the Apriori and FP-Growth algorithms to discover potential association patterns in engineering management. Figure 1 illustrates

Table 1. Training and Validation Scores of GBDT Model for Cost Prediction

Indicator	Training Set	Validation Set
MSE	3.28	3.51
RMSE	1.81	1.87

examples of high support and high confidence association rules discovered using FP-Growth.

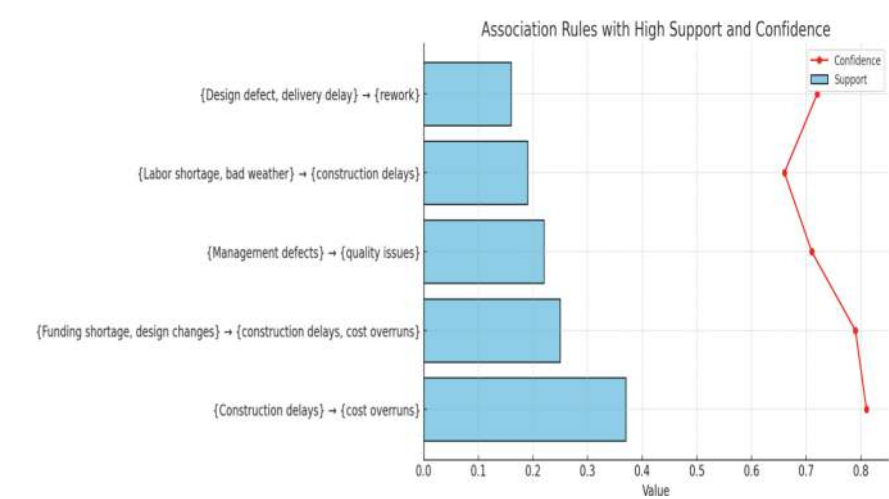


Fig. 1. Examples of High Support Association Rules

2.2 Model Training and Evaluation

After selecting suitable algorithms, extensive model training and evaluation were conducted. Taking classification tasks as an example, based on historical data of engineering projects, we constructed labeled training and test sets. Prior to training, data preprocessing steps such as standardization and imputation of missing values were performed, and feature engineering was carried out to remove features lacking predictive power. Subsequently, models were trained on the training set using 5-fold cross-validation, and their generalization performance was evaluated on the validation set [2]. Evaluation metrics included accuracy, precision, recall, and F1 score. Model parameters were also tuned to achieve optimal performance. Table 2 presents F1 scores of a random forest delay risk classifier under different parameter combinations. Ultimately, the optimal model parameters yielding the best performance were selected.

Table 2. Random Forest Model Parameter Tuning

Max Depth	Number of Estimators	Training Set F1	Validation Set F1
8	100	0.92	0.85
10	200	0.94	0.87
12	300	0.96	0.83

2.3 Model Optimization

After completing initial modeling, we continued to optimize the models to improve their performance and interpretability. For ensemble models, we conducted feature importance analysis, removing features with minimal contributions to the prediction target to simplify model complexity. Additionally, model blending was employed by combining weighted predictions from different models to achieve higher prediction accuracy. Furthermore, for black-box models, we utilized interpretability techniques such as SHAP (SHapley Additive exPlanations) to analyze the contribution of each feature to model predictions, enhancing model transparency [3]. Figure 2 illustrates the SHAP value distribution of different features in a cost prediction model.

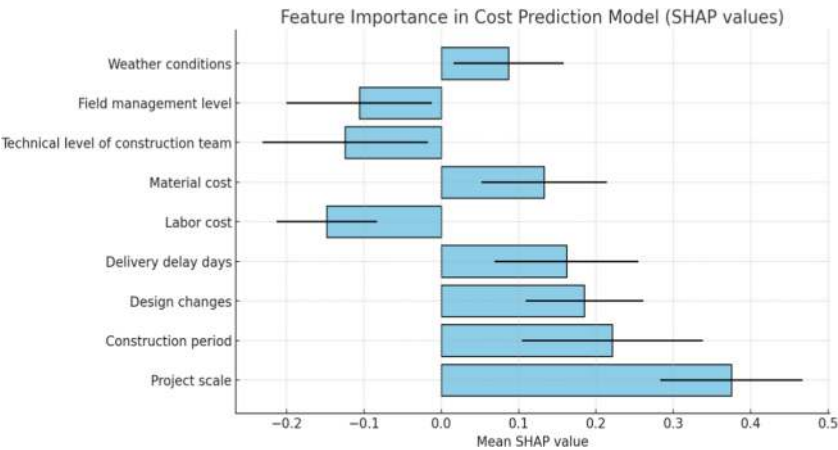


Fig. 2. Feature Importance in Cost Prediction Model (SHAP Values)

With these optimization measures, we increased the average prediction accuracy from an initial 82.6% to 87.3%, providing robust data support for project management decisions. Moreover, enhanced interpretability has made the decision-making process more transparent and rational.

3 System Implementation

3.1 System Architecture Design

To achieve efficient data processing and model deployment, we adopted a microservices architecture, dividing the system into multiple functional modules. The system primarily includes modules such as data collection, data management, model training, model serving, and visualization. The data collection module is responsible for gathering engineering project data from various sources such as progress reports, contract documents, and design drawings [4]. The data management module cleanses, stores, and manages the collected data. The model training module utilizes data mining algorithms developed earlier to train decision support models on processed data. The model serving module deploys trained models as services for use by other modules. The visualization module provides a user-friendly interactive interface to present analysis results and support configuration of decision parameters. Communication between modules is facilitated through REST APIs, ensuring a design that is highly cohesive and loosely coupled. Additionally, leveraging technologies like Kubernetes and Docker, we achieved scalable deployment of the system. During stress testing, a 100% increase in user traffic only required scaling up the replicas of model serving, ensuring stable response times [5]. Table 3 compares average response times under different levels of user traffic.

Table 3. System Response Time (ms) under Different User Traffic Levels

Concurrent Users	100	200	400	800	1600
Response Time (ms)	63	78	92	159	505

3.2 Data Storage and Management

In terms of data storage, we adopted a centralized data lake architecture managed through an object storage system. This approach offers enhanced scalability and cost efficiency compared to traditional databases. To improve data processing efficiency, we partitioned data based on project types, geographic regions, and other criteria. Additionally, for streamlined data management, tools like Hive were employed to automatically generate metadata and statistical information for data files stored in the object storage. This metadata service allows users to query data in distributed storage similar to querying a database [6]. Data quality was also prioritized significantly. We implemented automated processes for data quality detection and remediation, monitoring issues such as data missing or anomalies, and applying rule-based processes for remediation. Table 4 lists some common data quality rules.

3.3 Model Integration and Deployment

For ease of model integration and deployment, we adopted a microservices architecture based on Docker containers. Taking the cost prediction model as an example, we developed a RESTful API service using Flask:

Table 4. Common Data Quality Rules

Rule ID	Rule Description	Remediation Action
DQ_101	Progress payment amount is empty or less than 0	Fill with default value 0
DQ_202	Contract duration is empty or less than 7 days	Mark as anomalous
DQ_305	Project name is empty	Mark as anomalous, manual review

```
# Load the model
model = pickle.load(open('cost_model.pkl', 'rb'))
@app.route('/predict', methods=['POST'])
def predict():
    data = request.get_json()
    features = extract_features(data)
    prediction = model.predict(features)
    return jsonify({'cost': prediction.tolist()})
Subsequently, we packaged the service into a Docker image
and deployed and managed it using Kubernetes:
apiVersion: apps/v1
kind: Deployment
metadata:
  name: cost-model
spec:
  replicas: 3
  template:
    spec:
      containers:
        - name: cost-model
          image: cost-model:v1
---
apiVersion: v1
kind: Service
metadata:
  name: cost-model
spec:
  selector:
    app: cost-model
```

The above configuration creates a Deployment with 3 replicas and an associated Service, ensuring high availability access to the model service. Updating the model simply involves replacing the image version. This cloud-native approach enables flexible scalability of the system and simplifies the deployment process.

3.4 Visualization and Interactivity

To enhance the user-friendliness of the system, we developed an interactive web-based visualization interface. Users can intuitively understand various decision metrics such as project costs, progress, and risk levels through charts and reports. The system also allows users to set decision parameters and simulate outcomes under different decision scenarios. Figure 3 illustrates the workflow of the user interface [7]. Users first select the project and type of metrics for analysis; then, the system calls the corresponding model services to generate metric analysis results. Finally, results are presented in the form of charts or reports, with options for export and download. For instance, a typical cost prediction report includes comparisons between planned and actual costs over various periods, analysis of major cost drivers, and recommended optimization measures. Additionally, the visualization system provides rule-based alert functionalities. When anomalies or deviations from established control standards are detected, the system proactively alerts users, guiding timely attention to issues.

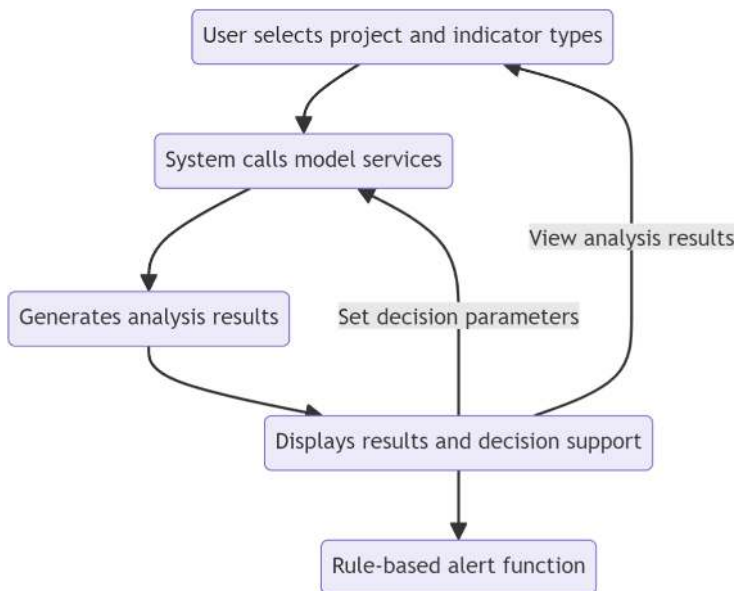


Fig. 3. Visualization and Interaction Workflow

4 Testing and Results

4.1 Test Environment and Data Preparation

To comprehensively validate the functionality and performance of our system, we constructed a test cluster simulating a production environment. The cluster consists of 10 physical machines, with 6 machines deployed as Kubernetes nodes, 3 dedicated to object storage services, and 1 serving as the cluster control plane. Each node is equipped with 32 GB of memory, an 8-core CPU, and 500 GB of disk space.

In terms of data, we extracted over 4000 samples from engineering project archives spanning nearly a decade, covering multiple domains such as construction, municipal services, and energy. These sample data include engineering contracts, change orders, progress reports, cost breakdowns, quality inspection records, among others, totaling 30 TB of data. We divided the data into training and test sets in an 8:2 ratio [8]. For key fields with missing data, such as contract duration and final completion dates, we supplemented the data by retrieving relevant additional documents or manually recording them by project managers. Additionally, we developed rule-based data cleansing tools to identify and rectify common data errors, such as incorrect formatting of amounts and reversed date sequences. Table 5 provides examples of some cleansing rules.

Table 5. Examples of Data Cleansing Rules

Rule ID	Rule Description	Action Taken
DC_01	Amount field contains non-numeric characters	Remove non-numeric characters
DC_05	Date field has illegal format	Standardize format based on context
DC_10	Start date is later than completion date	Flag as anomaly, manual review required

With the above data preparation, we have obtained a high-quality test dataset, laying the foundation for subsequent functionality and performance testing.

4.2 Functional Testing

In terms of functional testing, we designed a series of test cases covering core functionalities of the system, from data collection and model training to model deployment and interactive visualization. These tests were conducted in a simulated real-world testing environment.

Firstly, we tested the data collection module’s capability. Taking construction project construction logs as an example, we constructed samples in hundreds of document formats and verified that the system successfully parsed and stored over 99% of collected documents [9].

Next, functional testing was conducted on various decision support models. We randomly selected 100 samples from the test set for training, generated corresponding model instances, and deployed them. Subsequently, using these model services, we validated whether the analysis reports on the visualization interface were accurate and compared prediction results with actual known data. Table 6 presents the functional testing results of the cost prediction model.

Finally, we tested the functionality of various parameter interactions and decision simulation analysis on the interface. The results demonstrated that the system correctly responded to parameter adjustments and calculated new analysis values, providing decision-makers with visual decision support.

Table 6. Cost Prediction Model Functional Testing

Project ID	Actual Cost (million)	Predicted Cost (million)	Absolute Error (%)
P20301	18.27	17.93	1.86%
P19712	32.16	31.04	3.48%
P20105	55.38	52.19	5.75%
P18903	7.62	7.97	4.59%

4.3 Performance Testing

We conducted multidimensional load testing to assess system throughput, response times, and other performance metrics. Simulating access pressures from 100 to 5000 concurrent users, we recorded performance data such as response times and throughput rates. Figure 4 illustrates the relationship between concurrent users and average response times. It shows that with lower concurrent users (< 1000), response times remained under 100 ms, indicating sufficient computational resources to handle requests. However, as concurrent users increased, response times escalated rapidly.

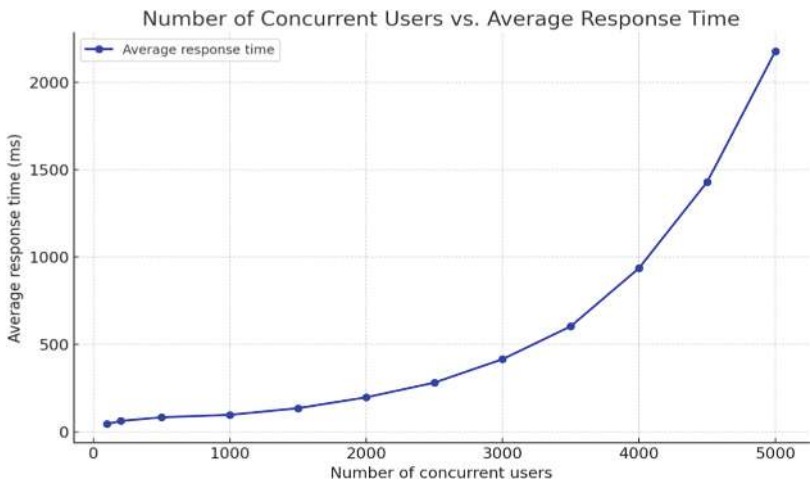


Fig. 4. Relationship between Concurrent Users and Response Time

To meet the demand for large-scale concurrent access, we optimized system resources. Initially, we increased the number of model service replicas and implemented request load balancing. Additionally, we scaled up data lakes and object storage to enhance I/O capabilities. Following these optimizations, the system achieved a maximum throughput of over 3500 requests per second, capable of handling access pressures from over 5000 concurrent users. Even under high loads, average response times were kept within 300ms, ensuring a favorable user experience. Table 7 outlines the key performance metrics of the optimized system.

Table 7. Performance Metrics of Optimized System

Metric	Value
Maximum Throughput	3583 req/s
Response Time at 5000 Concurrent Users	287 ms
Data Throughput	650 MB/s

4.4 Experimental Results

Combining the results of functional testing and performance testing, our Engineering Decision Support System has demonstrated excellent usability and scalability. The system has been piloted across multiple projects within our company, significantly enhancing decision accuracy and operational efficiency. For example, in Civil Engineering Project A, the system’s cost prediction and optimization analyses enabled rational resource allocation, resulting in a 10.7% cost savings. Additionally, leveraging the schedule delay risk model, we identified and mitigated potential delays in critical processes 35 days in advance, ensuring on-time project completion. In another construction project, Project B, early detection and correction of construction quality issues saved over 2.5 million RMB in rework costs. Moreover, the system uncovered underlying causes of quality issues through association rule mining, providing analytical basis for corrective actions. Project managers have highly praised the system: “Previously, decisions relied solely on experience. Now, with strong data support, judgments are more objective and scientific. Model analyses quantify risks and predict outcomes, bringing unprecedented transparency to the decision-making process.”

5 Conclusion

This paper introduced an Engineering Decision Support System based on data mining and machine learning technologies. By collecting and analyzing extensive project data, the system employs methods such as association rule mining, feature analysis, and predictive modeling to construct decision support models for cost control, schedule management, and quality assurance. Deployed with a microservices architecture, the system scales models as containerized services and offers a user-friendly interactive visualization interface for browsing predictive analysis reports and simulating decision parameters. Comprehensive functional and performance testing validated the system’s high availability and scalability. Real-world application projects have demonstrated significant improvements in decision quality and operational cost reduction. This system promotes the transformation of enterprise decision-making towards data-driven and model-based approaches, injecting new momentum into the high-quality development of the construction industry.

References

1. Wang, J.: Intelligent decision support system for building project management based on artificial intelligence. IOP Conf. Ser.: Earth Environ. Sci. **2665**(1), 012022 (2023)

2. Bilgin, G., Dikmen, I., Birgonul, M.T., Ozorhon, B.: A decision support system for project portfolio management in construction companies. *Int. J. Inf. Technol. Decis. Mak.* **22**(2), 705–735 (2023)
3. Zhang, S., Ou, W., Ren, G., Gao, Y., Cao, H.: Risk model and decision support system of state grid operation management based on big data. In: *International Conference on Smart Technologies and Systems for Internet of Things*, pp. 511–520. Springer, Singapore (2023)
4. Fröhlich, P., Mirnig, A.G., Falcioni, D., Weixelbaumer, B., Tscheligi, M.: Effects of reliability indicators on usage, acceptance and preference of predictive process management decision support systems. *Qual. User Exp.* **7**(1), 6 (2022)
5. El-Tourkey, M., Alshibani, A., Mohammed, A., Almuaini, Y., Alkubaisi, A.: An integrated decision support system for mobile crane selection. *Expert Syst. Appl.* **189**, 116124 (2022)
6. Zappone, A., Pavone, D., Natali, M., Schillaci, E., Braschi, C., Tozzi, M.: #920 A knowledge management system enabling telemedicine for hemodialysis: an ongoing project. *Nephrol. Dial. Transpl.* **39**(Supplement_1), gfac266.920 (2024)
7. Stankovics, P., Schillaci, C., Pump, J., Marucci, A.: A framework for co-designing decision-support systems for policy implementation: the LANDSUPPORT experience. *Land Degrad. Dev.* **35**(5), 1395–1408 (2024)
8. Wang, Y.: Environmental economic decision support information system based on GIS. In: *The International Conference on Cyber Security Intelligence and Analytics*, pp. 383–392. Springer, Cham (2022)
9. Neves-Silva, R., Camarinha-Matos, L.M.: Context-based decision support system for energy efficiency in industrial plants. *Sustainability* **14**(3), 1425 (2022)



Performance Analysis of Diversification Strategies in Internet Service Enterprises: An Empirical Study

Li Zhou^(✉)

School of Economics, Wuhan Business University, Hubei 430056, China
398749006@qq.com

Abstract. With the increasing uncertainty of the global economy, how to appropriately implement the diversified strategy has become a key issue that needs to be deeply analyzed. We collected 202 Internet Service Enterprises from Chinese A stock exchange during 2017 from 2021. The 1010 observation was collected for the research. Results show that there is a clear negative relationship between the diversified strategy of Chinese Internet Service Enterprises and their performance, which reveals that the diversified strategy may constitute a certain obstacle to the growth of enterprise performance. In order to more fully understand this phenomenon, research further introduced hefendal-herschmann index lag item, the regression analysis shows that both lag phase or lag phase ii, the influence of hefendal-herschmann index has not reached a significant level, the result further strengthens the diversified strategy failed to effectively improve the current and the future enterprise performance. In view of this, this study proposes that, for Internet Service Enterprises, a more prudent and prudent strategy should be adopted when implementing the diversified strategy. In view of the broad development prospects and fierce competition of the Internet service industry, enterprises should focus on their core business to consolidate their market position and ensure the steady growth of long-term competitiveness and performance.

Keywords: Internet Service Enterprises · Diversification · Diversified Strategy · Diversification Premium · Diversification Discount

1 Introduction and Literature

By December 2023, the number of Internet users in China had reached 1.92 billion, and the Internet penetration rate had reached 77.5 percent, according to the “*Statistical Report on China’s Internet Development*” released by the China Internet Network Information Center. The popularity of the Internet has increased the number of Internet Service Enterprises in China and they provided more abundant service products. The Internet services industry has shown a positive growth momentum in 2023. Among them, the Internet business revenue reached 1,748.3 billion yuan, and the software business revenue reached 1,232.8 billion yuan. The total profit of the Internet service industry maintained growth, among which, the total profit of the Internet reached 129.5 billion yuan, a slight

increase of 0.5%, and the total profit of the software industry reached 1.459.1 billion yuan, achieving an increase of 13.6%. In terms, information technology services, information security products, Internet life services and network sales are growing rapidly. In terms of regional distribution, the eastern and central regions have achieved rapid business revenue growth, especially in the Yangtze River Delta region, which leads in the Internet business revenue growth rate.

In the opportunity-filled area of Internet services, Whether the Internet Service Enterprises should take the diversified strategy or focus on the main business has become a dilemma facing the current development. Scholars at home and abroad have researched the effect of diversified strategy on performance through theoretical analysis and empirical research. Theoretically, if an enterprise concentrates the resources on its main business, it had better develop differentiated products and avoid taking risks in other businesses, thus dragging down the main business. However, enterprises can also obtain more market opportunities through diversified strategy, so as to disperse the main business risks and earn more stable profits. Therefore, whether the effect of diversified strategy on performance is positive or negative, is still unsettled. The conclusions drawn by a large number of empirical researches are often inconsistent.

Literature mainly includes three views: the first view is that the diversified strategy will inhibit the enterprise performance, and the inhibitory effect is heterogeneous. Zhao Shijun, Wang Yanyan, Yang Huihui (2022) [1] verified the inhibitory effect of diversified strategy on the enterprise performance through the data of high-tech enterprises, and they believed that the higher the uncertainty of the environment of the enterprise, the stronger the inhibitory effect will be. State-owned enterprises or enterprises with a higher competitive position in the industry will weaken this inhibitory effect. Zhang Bin, Yao Zhiyuan, Zhang Jiayi (2022) [2] considered that hiring specialized independent directors can effectively restrain the diversification discount of the enterprise. Wang Xiaoting (2019) [3] found that there is diversification discount phenomenon of home appliance enterprises. The negative relationship between diversification and performance of large home appliance enterprises is stronger than that of small home appliance enterprises. Jing Xinxin, Yang Fuming (2018) [4] found through clustering analysis that the diversified strategic reform of the listed real estate enterprises has different paths. More than 40% of the listed real estate enterprises are developing in the real estate sector, while 30% of the listed real estate enterprises are expanding to the non-real estate sector. In the short term, the diversified strategy in the real estate sector suppresses the enterprise performance. But the non-real estate diversified strategy has some lag effect on the enterprise performance.

The second view, contrary to the first, the diversified strategies would improve the enterprise performance. Zhang Liping, Fu Yumei (2019) [5] compared the business decisions of TBEA and Tianwei Baoding, and found that in the fiercely competitive market, the enterprise can achieve better performance by implementing the diversified strategy.

More scholars prefer to the third view, believing that diversification discount and diversification premium exist simultaneously. Whether the diversified strategy can improve the enterprise performance depends on the result of the interaction of diversification discount and diversification premium. Shen Yifei (2018) [6] considered the

relationship between the diversification of retail enterprises and the enterprise performance. Considering that the diversified strategy of retail enterprises can enjoy the advantage of sharing cost in the scope economy, the complementary advantage of synergistic effect resources, the advantage of entering high profit business of resource base and the cross-subsidy advantage of market forces, it will promote the performance growth of retail enterprises. But at the same time, the retail enterprise diversified strategy also has disadvantages. The disadvantage of the retail enterprises diversified strategy lies in the principal-agent problem. The diversified strategy of retail enterprises will increase such as the coordination cost, incentive degradation and bureaucratic distortion, and restrain the performance growth of retail enterprises. He researched with the data of listed retail enterprises, and found that the implementation of the diversified strategy will inhibit the growth of the performance of listed retail enterprises. Ren Li & Xiang Yu (2019) [7] believes that the diversification premium and the diversification discount of listed private enterprises coexists. The performance of listed private enterprises implementing the diversified strategy is more affected by the diversified strategy and the political and business relationship than the overall sample. Li Weiyao, Zhang Lijun, Wu Bing (2018) [8] explored the relationship between international diversified strategy and performance of Chinese enterprises, and believed that the performance of international diversification of state-owned enterprises is significantly better than that of private enterprises. Li Zhenqiu (2020) [9] believes that the implementation of professional strategy or related diversified strategy can improve the performance of retail enterprises. Further research shows that the implementation of related diversified strategy will promote the performance of enterprise. The implementation of unrelated diversified strategy will reduce the performance of enterprise. Li Juan, Ma Huimin (2017) [10] tested the difference between the interactive effect of industry diversification and internationalization on the enterprise performance in manufacturing and service enterprise groups. The research shows that the effect of group internationalization on the enterprise performance is U-shaped; the regulatory effect of industry diversification is positive before negative, and this adjustment effect is more obvious in manufacturing groups. Su Xin, Liu Haolong (2017) [11] believes that there is a negative correlation between the degree of industry diversification and the performance of listed companies, and there is a negative "U" relationship between product diversification and the performance of listed companies. Senior executive shareholding level has a positive regulating effect on the relationship between product diversification and performance; Beyond a certain threshold, senior executive shareholding level has a negative regulating effect on the relationship between product diversification and performance.

Research shows that different industries, different nature of enterprises, different relationship between government and business, and other factors will have an effect of enterprise diversified strategy. There are few literature on the effect of the diversified strategy on Internet Service Enterprises performance. For the special group of Internet Service Enterprises, there is still unconcern about whether it is the advantages or disadvantages of implementing diversified strategy of Internet Service Enterprises. Starting from micro enterprises, this paper studies the characteristics of listed Internet Service

Enterprises and verify the effect of the diversified strategy on Internet Service Enterprises performance, which has theoretical and practical significance for improving the performance of Internet Service Enterprises.

2 Research Methods and Hypotheses

From the perspective of the diversification premium and diversification discount theory, the performance of Internet Service Enterprises will be affected by the diversified strategy. Although the diversified strategy will disperse the resources of the main business, it will also obtain more market opportunities and disperse the market risks of the enterprise. It is uncertain whether the diversified strategy can effectively promote the performance of Internet Service Enterprises. Therefore, this paper only makes the assumption of relevance and drawn conclusions by empirical research. Hypotheses are proposed in this paper based on the literature review.

Hypothesis 1: Diversified strategy will have an effect on the enterprise performance.

Hypothesis 2: The Internet Service Enterprises diversification strategy may have a lag effect.

In order to testify to the effect of diversified strategy on the performance of Internet Service Enterprises, the control variables are introduced into the model while introducing diversification indicators. The specific model form is constructed as follows:

$$ARR_{it} = \beta_0 + \beta_1 HHI_{it} + \beta_2 control_{it} + \xi_{it} \quad (1)$$

Show in Table 1, among them, ARR_{it} is the depvar of the model, that is asset remuneration rate of the i enterprise in t year. The core indepvar is the Hufendal index of the i enterprise in t year (HHI_{it}). The logarithm of total assets, total assets turnover index, current ratio, cash ratio are the control variables choose according to the relevant literature. ξ_{it} is random error items.

3 Data Sources and Index Selection

According to the classification of national economy industry of enterprises, the Internet-related services and software & information technology services are the research objects. In this article the enterprises of Internet-related services and software & information technology services are defined as Internet Service Enterprises. The data were collected from the Chinese Internet Service Enterprises listing in A stock exchange during 2017–2021. With this data, we researched the effect of the diversified strategy on Chinese Internet Service Enterprises performance. The data were collected from listed enterprises, because it is more accurate and easier to obtain than that of non-listed enterprises. And, listed Internet Service Enterprises are relatively excellent enterprises in the Internet service industry, which can better represent the Internet Service Enterprises. Research the effect of the diversified strategy, it would provide a template for the development of Internet Service Enterprises. The Internet service industry is an emerging industry. A large number of Internet Service Enterprises do not with a long duration. The samples

Table 1. Data sources and Indicators selection

Indicator		Explanation
Depvar	ARR _{it}	ARR _{it} is the depvar of the model, That is asset remuneration rate of the i enterprise in t year
Core Indepvar	HHI _{it}	HHI _{it} is Hufendal index of the i enterprise in t year
control variable	LNTA _{it}	ln(total assets) is the natural logarithm of total assets. It is a control variable, which choose according to the relevant literature
	TAT _{it}	Total assets turnover is an indicator of the velocity of asset turnover. It is the control variables choose according to the relevant literature
	CRR _{it}	CRR _{it} is the ratio of current assets to current liabilities, which is used to measure the ability of current assets to be converted into cash before the maturity of short-term debt
	CAR _{it}	CAR _{it} is the ratio of the balance of the quick-moving assets after deducting the accounts receivable to the current liabilities, which can best reflect the ability of the enterprise to pay the current liabilities directly

selected are the enterprises listing in A stock exchange during 2017 to 2021 with a short duration is excluded. At last, 1010 observations were obtained, that is, 202 sample enterprises.

The data were collected from the original financial data of 202 Internet listed enterprises during 2017–2021 in CSMAR, and the data is true and reliable.

In this paper, the depvar was enterprise performance. The indicator to measure the enterprise performance generally refers to reflect the profitability of an enterprise. This paper, we selected the asset remuneration rate as the indicator to reflect the enterprise performance. The core indepvar is the diversification indicator. The Huffendal index (HHI_{it}) is selected to measure the degree of diversification. In this paper, HHI_{it} is calculated according to the classification of product revenue in the financial statements note database. The HHI_{it} measures the dispersion of the business income. The formula of HHI_{it} is:

$$HHI_{it} = 1 - \sum_{j=1}^n (X_{jt}/Y_t)^2 \tag{2}$$

Among which, X_{jt} represents the business income of the product j sales by the sample enterprise in year t, Y is the total revenue of the sample enterprise business in year t. The value of HHI_{it} is between 0–1. The larger the value of HHI_{it}, the more diversified the development strategy is, while the smaller the HHI_{it}, the more the company implements a professional strategy to produce. In addition, the The logarithm of total assets, total assets turnover index, current ratio, cash ratio were selected as control variables.

4 Descriptive Statistical Analysis

According to the diversified strategy classification method proposed by Wrigley, Specialization Rate (SR) and Correlation Rate (CR) were used to determine the diversified strategy type. The Specialization Rate index is measured by the proportion of products revenue. The Correlation Rate index is measured by the proportion of the largest business in the total revenue. According to the Specialization Rate, enterprises can be divided into four diversified strategy types: professional strategy, dominate diversified strategy, related diversified strategy and unrelated diversified. If the Specialization Rate (SR) of the enterprise is greater than 95%, it must be a professional enterprise. If $70\% < SR < 95\%$, it must be a dominate diversified enterprise. If SR is less than 70% and its business is related, it must be a related diversified enterprise; if SR is less than 70% and its business unrelated, it must be a unrelated business enterprises. Diversified strategy classification of Internet Service Enterprises annually is shown in Table 2.

Table 2. Annual statistics of diversified strategy classification

year	Professional strategy	Dominate diversified strategy	Related diversified strategy	Unrelated diversified strategy	Total
2017	6 (2.97%)	52 (25.74%)	94 (46.53%)	50 (24.75%)	202
2018	7 (3.47%)	55 (27.23%)	83 (41.09%)	57 (28.22%)	202
2019	8 (3.96%)	46 (22.77%)	92 (45.54%)	56 (27.72%)	202
2020	8 (3.96%)	55 (27.23%)	90 (44.55%)	49 (24.26%)	202
2021	8 (3.96%)	62 (30.69%)	81 (40.10%)	51 (25.25%)	202
Total	37 (3.66%)	270 (26.73%)	440 (43.56%)	263 (26.04%)	1,010

*The proportion of the total in parentheses

According to Table 1, it can be seen that Internet Service Enterprises choose the most related diversified enterprises, accounting for 43.56% of the samples. Secondly, they choose unrelated diversified strategy enterprises, accounting for 26.04% of the samples. This shows that Chinese Internet Service Enterprises want to diversify to spread market risks.

According to the samples, we get mean, variance and median of the ARR_{it} shown in Table 3.

It can be seen from the Table 3, the mean and median of ARR_{it} are the best performance of the related diversified enterprises; the variance shows the smallest variance of the related diversified enterprises, indicating that the related diversified enterprises have not only the higher performance but also the smaller difference between the related

Table 3. Descriptive statistics of ARR_{it}

year	Professional strategy			Dominate diversified strategy			Related diversified strategy			Unrelated diversified strategy		
	Mean	p50	SD	Mean	p50	SD	Mean	p50	SD	Mean	p50	SD
2017	0.08	0.06	0.09	0.06	0.07	0.07	0.06	0.06	0.11	0.05	0.05	0.06
2018	0.02	0.06	0.25	−0.03	0.03	0.23	0.03	0.05	0.10	−0.01	0.04	0.18
2019	0.04	0.06	0.16	0.02	0.04	0.21	0.03	0.05	0.13	−0.03	0.04	0.22
2020	0.00	0.00	0.10	0.00	0.04	0.16	0.02	0.05	0.19	0.02	0.05	0.11
2021	−0.01	−0.01	0.11	0.03	0.03	0.13	0.02	0.04	0.12	−0.08	0.02	0.52
Total	0.02	0.04	0.15	0.01	0.04	0.17	0.03	0.05	0.13	−0.01	0.04	0.27

diversified enterprises. In combination with the selection of related diversification enterprises in Table 2, it also shows that related diversified strategy is likely to be the most suitable for Internet Service Enterprises. The data show that most Internet Service Companies are adopting a diversified strategy, but is the degree of diversification appropriate? This paper will use the sample data to analyze the relationship between the degree of diversification and the enterprise performance.

5 Regression and Outcome Analysis

A panel regression analysis of the model was performed using the Stata17, and the regression results are shown in Table 4. In Table 4, column (1) is fixed effect model and column (2) is random effect model. According to Table 4, the diversification index (HHI) has a negative impact on the performance of Internet Service Enterprises, whether the fixed effect model or the random effect model is selected. Using the hausman-test, we selected a fixed-effects model for the analysis. Column (3) and column (4) in Table 4 are fixed and random effects models with annual control variables, respectively. By hausman test, we selected the fixed effects model column (3) for analysis. According to column (3), we found that the diversification index (HHI) has a significant negative impact on the performance of Internet Service Enterprises, that is, the diversification will inhibit the performance of Internet Service Enterprises. Compared with column (1), the regression coefficient of HHI after adding the annual control variable is still significantly negative, and the absolute value of the coefficient is even larger. This indicates that taking the annual control variable the in the regression, the impact of HHI on the enterprise performance is greater. Hypothesis 1 holds.

To test Hypothesis 2, the lag term of HHI was added to the regression. In Table 4, the lagging first phase of the HHI was added is to column(5), and the lagging II phase of HHI was added to column (6). We can see that the lagging phases of HHI are not significant, indicating that diversified strategy will not only inhibit the enterprise performance in the current period, but also will not have a positive impact on enterprise performance in the next one or two years. Hypothesis 2 does not hold. This is what distinguishes the Internet

service industry from some other industries. Although the adoption of diversified strategy in some industries has a significant negative impact on the current performance, with the passage of time, the accumulation of business outside the main business will begin to contribute to the enterprise performance to a certain extent. This may be related to the characteristics of the Internet service industry. The knowledge of the Internet service industry updates quickly, and the accumulation of experience may not bring more help.

Table 4. Regression analysis

depvar		Enterprise performance					
		(1)	(2)	(3)	(4)	(5)	(6)
explanatory variable	HHI _{it}	−0.1086* (−1.82)	−0.0188 (−0.50)	−0.1450** (−2.45)	−0.0250 (−0.66)	−0.1344* (−1.70)	−0.2522** (−2.37)
control variable	LNTA _{it}	0.1178*** (5.95)	0.0335*** (4.25)	0.1524*** (7.43)	0.0367*** (4.61)	0.1180*** (4.16)	0.1676*** (3.79)
	CRR _{it}	0.0109 (1.63)	0.0120** (2.37)	0.0088 (1.34)	0.0117** (2.32)	0.0129 (1.38)	0.0130 (1.11)
	CAR _{it}	0.0034 (0.36)	0.0011 (0.15)	0.0002 (0.02)	−0.0003 (−0.04)	0.0055 (0.42)	−0.0015 (−0.09)
	TAT _{it}	0.0345 (1.55)	0.0193 (1.62)	0.0485** (2.20)	0.0199* (1.67)	0.0460* (1.81)	0.0464 (1.54)
L.HHI _{it}						−0.0325 (−0.42)	−0.0469 (−0.45)
L2.HHI _{it}							0.0750 (0.67)
Annual effect				✓	✓		
constant		−2.5374*** (−5.91)	−0.7471*** (−4.34)	−3.2099*** (−7.27)	−0.7708*** (−4.45)	−2.5375*** (−4.15)	−3.5860*** (−3.76)
N		1010	1010	1010	1010	808	606
r _{2_w}		0.0490	0.0311	0.0877	0.0524	0.0394	0.0532
Hausman test		The fixed effects model outperforms the random effect model					

Note: ***, ** and * are significant at 1%, 5% and 10%, respectively, the t value based on the White in brackets

6 Conclusions and Recommendations

In this article, 202 samples were selected from the Chinese Internet Service Enterprises listing in A stock exchanges during 2017–2021. The data were analyzed to research the effect of diversified strategy on the performance of Internet Service Enterprises. The result shows that: the diversified strategy of Chinese Internet Service Enterprises can significantly negatively related with the performance of the company.

The research results show that the degree of diversification of Chinese Internet Service Enterprises is significant, that is, the higher the degree of diversification of the company, the worse their performance. This shows that the diversified strategy of Internet Service Enterprises will restrain their performance improvement. After adding the

lag term of HHI, we found that the first and second phases of HHI were not significant, indicating that the diversified strategy will not only inhibit the current enterprise performance, but also cannot have a significant impact on the future enterprise performance. This suggests that Internet Service Enterprises must be more cautious about their diversification. The Internet service industry has a broad market development prospect, and enterprises should focus more on their main business.

Of course, the research in this paper has certain limitations. At present, the sample size of listed Internet Service Enterprises in China is relatively small. With the continuous expansion of the sample size of Chinese Internet Service Enterprises, the future research data on Internet Service Enterprises will be more sufficient, which can better reveal their development laws and guide the development of China's Internet Service Enterprises.

Acknowledgements. [Fund Project] Collaborative Education Project of Ministry of Education "Smart Business and Big Data Campus Training Base of Wuhan Business School" (231106396213440).

References

1. Zhao, S., Wang, Y., Yang, H.: Research on the Effects of diversified operation on high-tech enterprise performance. *Commun. Finan. Account.* **23**, 54–59 (2022)
2. Zhang, B., Yao, Z., Zhang, J.: Diversified operation, industry specialty of independent directors and enterprise performance——based on A-share manufacturing listed companies. *Friends Account.* **20**, 66–71 (2022)
3. Wang, X.: Analysis of the relationship between diversified operation, customer concentration degree and enterprise performance. *J. Commer. Econ.* **20**, 127–130 (2019)
4. Jing, X., Yang, F.: Does the diversified strategic change promote or restrain enterprises' performance——the empirical analysis based on china's listed real estate enterprises. *China Bus. Market* **32**(09), 65–73 (2018)
5. Zhang, L., Yumei, F.: Does diversification reduce corporate performance?——Case study based on TBEA and Tianwei Baoding. *Friends Account.* **06**, 132–136 (2019)
6. Shen, Y.: Selection of diversification and normalization strategy of retail enterprises in China——Based on the empirical research of listed retail enterprises. *J. Commer. Econ.* **12**, 132–135 (2018)
7. Ren, L., Xiang, Y.: Research on the relationship between government and business, diversification and enterprise performance of Chinese private enterprises. *Fujian Tribune* **08**, 67–77 (2019)
8. Li, W., Zhang, L., Bing, W.: International diversification and enterprise performance: a four-stage inverted M-model. *Sci. Res. Manag.* **39**(06), 57–66 (2018)
9. Li, Z.: Shareholding structure, diversified strategy types and business performance of listed retail enterprises. *J. Commer. Econ.* **23**, 114–116 (2020)
10. Li, J., Ma, H.: Research on the relationship between group corp internationalization, industry diversification and enterprise performance. *Commun. Finan. Account.* **24**, 21–25 (2017)
11. Xin, S., Liu, H.: The effects of diversification on corporate performance: the moderating role of senior managerial ownership. *On Econ. Prob.* **04**, 100–107 (2017)

Data Governance and Ethics



Research on the Impact and Innovation Strategies of Enterprise Digital Transformation on Energy Utilization Efficiency

Juanjuan Tian and Ruiqin Tang^(✉)

School of Management, Wuhan Donghu College, Wuhan, China
trq_dhxy@163.com

Abstract. With the rapid development of the socio-economy, corporate digital transformation has become an essential pathway for fostering economic growth and enhancing competitiveness. This paper first analyzes the impact of digital transformation on the production and operational management models of enterprises, highlighting that the application of digital technologies can enhance energy utilization efficiency and reduce resource waste. It addresses the challenges and issues arising in corporate digital transformation, such as information security risks and technological costs, and discusses innovative strategies that should be implemented during digital transformation, such as strengthening data security and enhancing employees' digital literacy. Finally, through the discussion of actual cases, the paper substantiates the positive effects of digital transformation on energy utilization efficiency and underscores the importance of these innovative strategies for corporate energy conservation, emissions reduction, and sustainable development. Therefore, during the digital transformation, enterprises should focus on improving energy utilization efficiency and actively explore innovative strategies that are compatible with their own development, to achieve a win-win situation in terms of economic and social benefits.

Keywords: Corporate Digital Transformation · Energy Utilization Efficiency · Innovative Strategies · Production and Operational Management · Sustainable Development

1 Introduction

With the rapid development and application of information technology, corporate digital transformation (hereinafter referred to as DT) has become a significant trend in today's socio-economic development. This significant change directly alters the traditional models of production and operational management, bringing unprecedented opportunities and challenges to enterprises. In this context, the efficiency of energy use, as a crucial factor for sustainable development, has garnered increasing attention [1]. The impact of corporate DT on energy utilization efficiency (hereinafter referred to as EUE) is of great interest, as the widespread application of digital technologies can effectively improve EUE in production and operation processes, reduce costs, and minimize resource waste.

DT not only improves the internal production and operational models of enterprises but also provides more precise data on energy use, enabling enterprises to manage energy resources more effectively and achieve their goals of energy conservation and emission reduction. However, DT also introduces a series of new challenges and issues. During the process of DT, enterprises may face risks related to information security, high technological costs, and other issues that directly affect the improvement of EUE. Therefore, enterprises need to develop corresponding innovative strategies to better address the challenges encountered during DT.

This paper aims to explore the significant role of DT in sustainable development by conducting an in-depth study on the impact of corporate DT on EUE and the corresponding innovative strategies. Through case analysis and theoretical research, this study hopes to reveal how DT can promote enterprises to enhance EUE and play a positive role in energy conservation, emission reduction, and sustainable development. Furthermore, this research will provide references and guidance for future DTs of enterprises, laying a solid foundation for enterprises to achieve a win-win situation in economic and social benefits through DT. By providing a systematic study that delves deep into the impact of corporate DT on EUE and innovative strategies, this paper aims to offer theoretical support and practical guidance for integrating corporate DT with the enhancement of EUE [2].

2 Literature Review and Theoretical Foundations

2.1 Theoretical Framework of Corporate DT

Corporate DT is one of the hot topics in the socio-economic development today, involving the comprehensive application of information technology, digital technology, and data analysis in enterprises to achieve business optimization, innovative development, and enhanced market competitiveness. In the process of DT, enterprises need to integrate resources in technology, management, and organizational structure to transform traditional business models into digital and intelligent ones, thereby improving operational efficiency and profitability. When researching the theoretical framework of corporate DT, it is beneficial to refer to existing research achievements and combine them with practical case analysis. Digital strategy theory is one of the important theoretical bases of corporate DT. According to Porter and Millar (1985), enterprises need to integrate digital technologies into their strategies to innovate and enhance business processes. The Technology-Organization-Environment (TOE) framework can serve as a theoretical basis for analyzing corporate DT, emphasizing the interplay among technology, organization, and environment, and the success of DT is closely related to the coordinated operation of these three elements. Innovation theory is an important theoretical support for corporate DT. According to Schumpeter's theory of innovation, innovation drives enterprise development, and DT is an innovative approach that benefits the enhancement of competitiveness and market position. Resource-based theory also plays a significant role in the theoretical framework, as DT enables enterprises to maximize the use and integration of internal and external resources, improving production efficiency and innovation capacity, laying the foundation for sustainable development. Therefore, corporate DT can be regarded as a comprehensive strategic decision-making process

involving factors such as technology, organization, and management. In practical applications, enterprises need to formulate corresponding digital strategies and innovative strategies based on their characteristics and development needs, with appropriate theoretical support, to achieve the goals of optimizing EUE and enhancing economic benefits, as illustrated in Fig. 1.

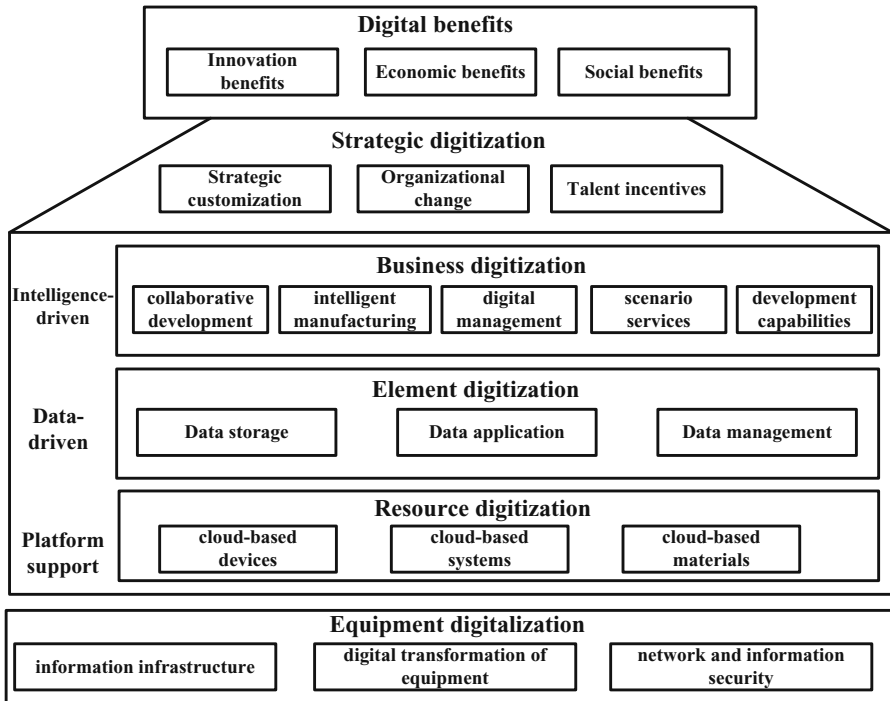


Fig. 1. Framework for enterprise DT.

2.2 Relevant Theories on EUE

EUE refers to the utilization efficiency of energy resources by enterprises in the production and operation process, which will reflect whether the enterprise is rational in resource utilization [3]. It is of great significance to improve the economic benefits of enterprises, reduce resource consumption, and achieve sustainable development. When studying the impact of DT of enterprises on EUE, the construction of relevant theories and models is crucial. According to the resource conservation theory, which is an important theoretical basis for EUE research, enterprises can reduce resource consumption costs, improve production efficiency, and achieve sustainable development by rational utilization and management of resources. In DT, energy resources are one of the important production factors for enterprises. The application of digital technology helps enterprises manage and utilize energy resources more efficiently, and improve EUE.

Environmental management theory also plays an important role in the research of EUE. It emphasizes that enterprises should pay attention to environmental protection and resource conservation in the production and operation process, adopt corresponding strategies and technical means, and achieve a win-win situation of efficiency and environmental protection. For DT, it can help enterprises monitor and manage energy utilization, thereby discovering and solving the problem of energy waste and improving the efficiency of energy resource utilization.

The theory of energy conservation and emission reduction also plays an important guiding role in improving the EUE of enterprises. With the enhancement of environmental protection awareness and the improvement of international energy conservation and emission reduction requirements, enterprises need to take effective measures to reduce energy consumption and carbon emissions. DT provides an important way for enterprises to achieve energy conservation and emission reduction goals, but also provides certain opportunities. To ensure the improvement of EUE, it is necessary to combine the influence and synergy of multiple factors, and through the application of digital technology, data analysis and other technical means, make energy management more intelligent and efficient. Therefore, when studying the impact of enterprise DT on EUE, it is necessary to deeply understand the relevant theories of resource conservation, environmental management, and energy conservation and emission reduction, in order to build a scientific and effective research framework, Promote the improvement of EUE and the realization of sustainable development in enterprises [4].

2.3 Research Progress on the Relationship Between DT and EUE

Previous studies have revealed the close relationship between DT and EUE, providing a rich research foundation for in-depth exploration of this topic. According to some studies, DT can significantly improve the EUE of enterprises. The application of digital technology enables enterprises to monitor and manage energy consumption more finely. Data analysis has the potential to identify energy-saving opportunities, optimize production processes and equipment utilization, thereby reducing energy consumption costs and improving EUE. Through DT, enterprises can achieve effective allocation and utilization of resources, promoting the rationalization of energy consumption. At the same time, the improvement of EUE through DT of enterprises depends on the formulation and implementation of innovative strategies. Although DT provides technical means and data support for enterprises, only by combining innovative strategies can digital technology be applied to energy management practices, ultimately achieving the best results. Innovative strategies involve the reform of internal management systems, employee skill training, technological updates, and other aspects, promoting the continuous improvement of EUE. The impact of DT on EUE is influenced by internal and external environmental factors of enterprises. When conducting DT, enterprises need to fully consider the synergistic effect of multiple factors such as technology, organization, policies, etc. These factors will directly or indirectly affect the EUE after DT. After fully studying environmental factors, corresponding strategies should be formulated to promote practice. The relationship between DT and EUE is very complex, requiring comprehensive consideration of various factors such as technology, management, and environment. Research on the impact of DT on EUE, as well as the formulation and

implementation of innovative strategies, can provide theoretical guidance and practical support for enterprises to improve EUE and achieve sustainable development. In order to promote DT, Chinese energy enterprises adhere to strategic leadership [5]. Table 1 compares the DT strategies of some energy companies, which based on their own development foundation and business areas, have formulated long-term plans and system designs for the development goals and implementation paths of DT.

Table 1. Overview of DT Strategies for Some Energy Enterprises.

Company Name	Strategic Planning	Development Goals	Implementation Path
National Energy Group	“National Energy Group DT Strategy”, “National Energy Group Network Security and Informatization 14th Five-Year Overall Plan”	“Platform-based Development, Digital Operation, Ecological Collaboration, Industrial Chain Collaboration, Intelligent Production”	Building intelligent production, smart management, smart operation system
State Grid Corporation of China	“DT Development Strategy Outline”	“Equipment, Operations, Management, Collaboration, Digitalization”,	Adhere to the development layout of “One Core Four Wings”, iterate and improve the business middle platform, build PM3.0 management system
Southern Power Grid Corporation	DT and Digital Power Grid Construction Action Plan	Strategic orientation: transformation towards “digital power grid operators, energy industry value chain integrators, energy ecosystem service providers”	First proposed the concept of “digital power grid”, 4321 architecture for building digital power grid

(continued)

Table 1. (continued)

Company Name	Strategic Planning	Development Goals	Implementation Path
China Datang Corporation	3549 DT strategy	Three major improvement directions: “Group control, operation production, innovation development”; Five major improvement goals: “new positioning, new control, new capabilities, new architecture”	implement four major projects and coordinate the construction of nine major digital architecture platforms
China Petrochemical Group	“14th Five-Year” Informatization Development Plan and DT Strategy	with the core of improving the digitalization level of the industry	Promote the construction of “Four Clouds”, “Three Systems”, “Two Platforms” in the mode of “data + platform + application”

3 Analysis of the Impact of DT on EUE

3.1 Core Elements of DT

DT, as a crucial process for enterprises to achieve informatization and intelligentization, relies on the synergistic effect of multiple core elements. To study the impact of corporate DT on EUE, it is essential to analyze these core elements to uncover the mechanisms through which they affect energy efficiency. Technological innovation is one of the key elements of DT, including the application and integration of hardware, software, cloud computing, and big data analytics. Enterprises enhance their energy management level, reduce energy waste, and improve efficiency by incorporating advanced energy management systems, intelligent devices, and sensors to enable real-time monitoring, data analysis, and intelligent control [6].

Organizational structure and cultural transformation are indispensable core elements of DT. Enterprises need to modify traditional organizational structures and management models, promote information sharing, collaborative innovation, and cross-departmental cooperation, and facilitate the comprehensive implementation and application of digital technologies. Establishing a data-centric corporate culture encourages employee participation in DT, pushing forward the renewal and transformation of energy management concepts.

Talent development and capacity building are one of the pivotal elements of DT. Enterprises require a team proficient in the skills and expertise needed for DT, providing

relevant training and technical support to employees, fostering their enthusiasm and creativity for participating in DT. High-quality talent not only secures the process of DT but also plays a crucial role in enhancing EUE.

Data security and privacy protection are also core elements of DT. As the level of corporate digitization increases, issues related to data security become more pronounced. Enterprises need to establish a comprehensive data security system to ensure the security and reliability of energy utilization data while complying with relevant laws and regulations to protect user privacy and ensure the smooth progression of DT [7].

The core elements of DT, including technological innovation, organizational structure and cultural change, talent development and capacity building, and data security and privacy protection, work together to promote the enhancement of EUE and help achieve sustainable development goals.

3.2 The Direct Impact of DT on Energy Efficiency of Enterprises

The direct impact of enterprise DT on EUE is multifaceted. DT enhances the ability to collect and monitor energy data, enabling enterprises to achieve real-time monitoring, data collection, and analysis of energy consumption, thereby gaining a deeper understanding of energy consumption patterns and patterns. This helps identify energy waste issues, adjust production plans in a timely manner, optimize production processes, and ensure the efficiency and rationality of energy utilization. In addition, DT improves the refinement level of energy management, providing enterprises with diverse energy management tools and platforms, which can achieve online monitoring and intelligent control of energy consumption equipment. These tools and platforms enable enterprises to analyze and optimize energy consumption based on real-time data, and formulate more accurate energy management strategies, So as to reduce energy consumption costs and improve EUE. DT also promotes the conservation and utilization of energy resources. Through the application of digital technology, enterprises can optimize the allocation and utilization of energy resources, achieve equipment automation control and intelligent operation, improve production efficiency while reducing energy waste. DT enables enterprises to adjust energy use plans more flexibly, achieve on-site energy conservation and optimal utilization of resources. It promotes energy innovation and sustainable development of enterprises. With the support of digital technology, enterprises can carry out innovative exploration of energy management, promote the application of clean energy and the development of renewable energy, improve EUE while reducing dependence on traditional energy, and achieve sustainable development goals. To examine the relationship between DT and enterprise EUE, this article constructs the following benchmark model:

$$TFP_{i,t} = \beta_0 + \beta_1 DT_{i,t-1} + Controls_{i,t} + Firm_i + Year_t + \varepsilon_{i,t} \quad (1)$$

In the model, energy enterprise utilization efficiency (*TFP*) serves as the dependent variable, while *DT* acts as the core explanatory variable. Represents the aforementioned control variables, and denotes the random error term. Considering that there is a certain time lag in the impact of *DT* on the EUE of enterprises, a one-period lag is applied to the *DT* variable. This not only takes into account the transmission time between variables in practice but also mitigates endogeneity problems caused by reverse causality.

3.3 Application Framework of Digital Technology in Energy Management

The application cases of digital technology in energy management are rich and diverse, among which some technologies are widely used in enterprise DT to improve EUE. Data analysis technology plays a role in energy consumption prediction and optimization. Enterprises can use data analysis technology to analyze historical energy consumption data, accurately identify the patterns and trends of energy consumption, predict future consumption situations, optimize production plans and equipment scheduling, arrange energy use reasonably, reduce peak electricity consumption, and ultimately improve EUE; Intelligent control technology can monitor the operating status and energy consumption of equipment in real time, automatically adjust equipment operating parameters based on data feedback, and achieve dynamic optimization of energy consumption; Optimization technology can comprehensively consider multiple factors and analyze the optimal energy utilization plan through mathematical models [8]. Enterprises can develop reasonable energy consumption strategies based on the results of optimization technology, improve efficiency, and reduce costs; The application of machine learning technology in energy management is becoming increasingly widespread. Through machine learning algorithms, enterprises can learn energy consumption patterns and patterns from a large amount of data, identify potential energy-saving opportunities, automatically adjust energy management strategies, continuously optimize energy management processes, and improve EUE. The application of these technologies helps enterprises achieve intelligent and refined energy management, improve EUE, and promote the achievement of sustainable development goals. The application framework of digital technology in energy management is shown in Fig. 2.

3.4 Data Analysis of the Impact of DT on EUE

In order to conduct in-depth research on the impact of enterprise DT on EUE, it is necessary to conduct strict and detailed data analysis. Through data analysis, the specific impact mechanism of DT on EUE is revealed, providing a basis for formulating innovative strategies. Starting from the data analysis of energy consumption, a comparative analysis of energy consumption data before and after enterprise DT can be conducted to discover the impact of DT on EUE, such as whether enterprise energy consumption has decreased after DT, whether EUE has been improved, etc. Based on the detailed analysis of energy consumption data, the actual impact of DT on EUE can be evaluated. Trend analysis of energy consumption data is also essential [9]. Through time series analysis, we can observe the changes in energy consumption after the implementation of DT, such as whether there are seasonal fluctuations in energy consumption and whether there is a steady improvement in EUE before and after DT. Trend analysis can help us better understand the long-term impact of DT on enterprise EUE. Correlation analysis can explore the correlation between DT and EUE. By establishing a correlation model between energy consumption data and DT related indicators, the correlation between the two can be found. This can evaluate the impact of different aspects of DT on EUE, provide data support for formulating specific innovation strategies. Data mining technology can mine hidden information in the data, discover potential optimization opportunities, such as using data mining technology to discover abnormal patterns of energy consumption,

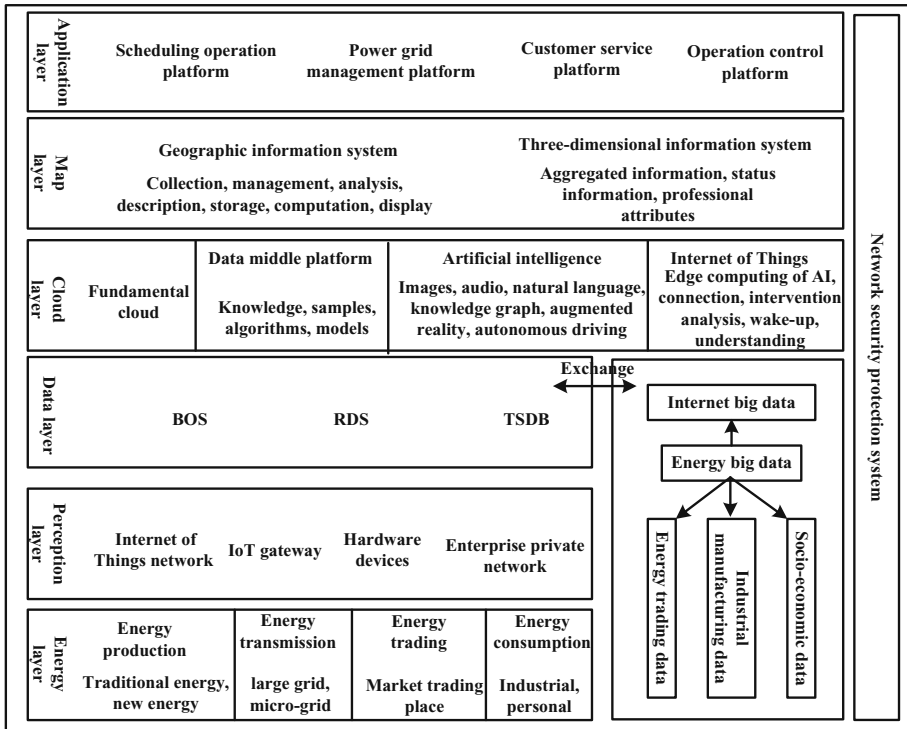


Fig. 2. Application framework of digital technology in energy management.

identify potential causes of energy waste, provide accurate energy management suggestions for enterprises, and formulate corresponding innovation strategies to improve EUE. By combining these analysis methods, we can deeply explore the impact of DT on the EUE of enterprises, evaluate the actual effects of DT, and provide scientific basis for enterprises to formulate innovation strategies.

4 Innovative Strategies for EUE in the DT of Enterprises

4.1 Algorithm Application for Optimizing Energy Management

Enterprises can achieve refinement of energy management through intelligent optimization algorithms to improve EUE. This algorithm can automatically optimize calculations based on real-time energy consumption data and production demand, find the best energy use plan, and machine learning algorithms also play an important role in optimizing energy management. By training historical data, machine learning algorithms can discover patterns and patterns of energy consumption, predict future energy consumption trends, and propose corresponding optimization suggestions. The application of data mining algorithms can help enterprises discover optimization opportunities in energy management and provide a basis for formulating innovative energy management strategies. The application of predictive algorithms in energy management is equally

crucial, as it can help enterprises understand future energy demand in advance, formulate corresponding energy supply plans, and avoid energy shortages and waste. Through the application of intelligent optimization algorithms, machine learning algorithms, data mining algorithms, and prediction algorithms, enterprises can achieve refined and intelligent energy management, improve EUE, reduce costs, and achieve sustainable development goals. The application of these algorithms will drive innovation and improvement in enterprise energy management, laying a solid foundation for achieving sustainable development [10].

4.2 EUE Improvement Strategies Driven by Data

By using data-driven methods, enterprises can more accurately identify potential energy-saving opportunities and optimization space. Through big data analysis, enterprises can achieve refined management of energy consumption, and further understand the energy consumption situation in each link. Based on real-time energy consumption data collection, organization, and analysis, enterprises can discover peak periods of energy consumption and high energy consuming equipment, propose corresponding energy management plans, and achieve energy-saving and emission reduction goals. Data driven predictive analysis also plays an important role in improving EUE. Through predictive analysis technology, enterprises can predict future energy demand based on historical data and trends, formulate reasonable energy use plans, achieve optimization and conservation. Data driven automation control systems are one of the key strategies to improve EUE. Through intelligent control systems, real-time monitoring and control of equipment can be achieved, and equipment parameters can be automatically adjusted to improve EUE. The application of data mining technology can also explore potential problems and optimization opportunities, accurately identify bottlenecks and optimization points, and propose effective improvement plans. The data-driven strategy for improving EUE is an important innovative strategy in enterprise DT, which can achieve intelligent and refined energy management, reduce costs, and promote the improvement of energy management level [11].

4.3 Future Oriented Innovative Technologies for EUE

Future oriented innovative technologies for EUE will become the key to achieving sustainable development for enterprises, and renewable energy technology will become an important direction for future EUE. Enterprises can achieve diversified energy supply and reduce dependence on traditional fossil fuels by integrating renewable energy technologies such as solar energy, wind energy, and hydropower. The application of digital technology can achieve real-time monitoring and management of renewable energy generation, improve utilization efficiency, and reduce energy costs.

The intelligent energy management system will be one of the important paths to improve EUE in the future. By integrating IoT, big data analysis, and artificial intelligence technology, enterprises can achieve real-time monitoring and intelligent regulation of energy equipment. Intelligent energy management systems can optimize equipment operating parameters, adjust energy usage plans, allocate energy resources reasonably based on real-time needs, and achieve optimal energy consumption. The development

of energy storage technology will also play a crucial role in improving future EUE. By utilizing advanced energy storage technologies such as battery storage and thermal storage, enterprises can solve the problem of high volatility in renewable energy and achieve stable energy supply. The application of digital technology can achieve remote monitoring and intelligent control of energy storage equipment, improve EUE, and optimize energy consumption patterns [12].

Ecological design and green building technology are also important directions for future EUE innovation. By utilizing ecological design principles and green building technologies, enterprises can reduce building energy consumption and improve indoor environmental comfort, thereby reducing the consumption of traditional energy. The application of digital technology can help enterprises achieve real-time monitoring and adjustment of building energy efficiency, improve overall EUE, and provide broader development space for enterprise DT. Through the application of renewable energy technology, intelligent energy management systems, energy storage technology, and ecological design technology, enterprises can achieve more efficient energy utilization, reduce energy waste, reduce environmental impact, and achieve sustainable development goals.

5 Conclusion

In this study, the positive impacts of corporate DT on EUE, as well as the challenges and opportunities arising during the DT process, have been analyzed. By exploring innovative technologies for future-oriented energy efficiency and the application of algorithms to optimize energy management, the paper provides strategic recommendations for enterprises undergoing DT. Corporate DT has significant potential and advantages in enhancing EUE. The widespread application of digital technologies enables real-time monitoring and value extraction of energy data, playing a key role in improving energy efficiency and reducing energy consumption. DT can not only optimize production processes and enhance energy efficiency but also bring more business opportunities and competitive advantages to enterprises.

DT also faces challenges and obstacles, such as technological investment, talent development, and security risks. However, it is these challenges that spur enterprises to further explore and innovate in DT, creating more opportunities for development. The importance of innovative technologies in EUE within the context of DT is emphasized for the future. Through the application of renewable energy technologies, smart energy management systems, energy storage technologies, and ecological design technologies, enterprises can achieve more efficient methods of energy utilization, reduce energy consumption costs, and promote environmentally sustainable development. Further inquiry into the optimization of energy management algorithms, such as the application of intelligent optimization algorithms, machine learning algorithms, data mining techniques, and predictive analytics, constitutes an important strategy for enterprises to enhance EUE in the future. The study of the impacts of corporate DT on EUE and its innovative strategies holds significant importance. Through the application of digital technologies and innovative strategies, enterprises can achieve more efficient energy use, enhance competitiveness, and realize sustainable development.

Acknowledgment. This work was supported by Scientific Research Plan Project of the Hubei Provincial Department of Education in 2022 “Research on the Impact of Digital Transformation on Improving Enterprise Energy Efficiency” (Project No.:B2022318).

References

1. Shi, R.: Exploration of digital transformation of new energy enterprises based on ERP concept. *China Arab Sci. Technol. Forum* (in Chinese and English) **30**(3), 54–57 (2024)
2. Ren, Y., Tian, Z., Liu, C., et al.: The impact of digital transformation on total factor productivity of enterprises: based on empirical evidence from Chinese listed energy enterprises. *Res. Technol. Econ. Manag.* **20**(1), 96–101 (2024)
3. Ren, Y., Tian, Z., Liu, C., et al.: How does digital transformation affect the performance of private energy enterprises—Empirical evidence from text recognition of annual reports of Chinese listed companies. *Tech. Econ.* **43**(1), 41–54 (2024)
4. Liu, W., Song, L.: The impact of enterprise digital transformation on energy utilization efficiency: based on the regulatory role of management power. *J. Shanghai Inst. Electr. Eng.* **26**(6), 361–366 (2023)
5. Ren, Y., Tian, Z., Liu, C., et al.: Digital transformation and green innovation in energy enterprises: a study on the heterogeneity of production processes and business models. *Enterpr. Econ.* **42**(12), 96–106 (2023)
6. Feng, L., Yang, K., Cui, J.: Research on the Impact of digital transformation on investment efficiency of traditional energy enterprises. *Coal Econ. Res.* **43**(10), 80–86 (2023)
7. Fucai, L., Qin, Y., Yuanbin, X.: The impact of enterprise digital transformation on total factor energy efficiency: from the perspective of specialized division of labor. *Contemp. Finan.* **23**(11), 3–15 (2023)
8. Ma, Z., Liu, P., Cai, L.: Has digital transformation improved green all factor energy efficiency—Empirical evidence from Chinese A-share industrial listed companies. *Enterpr. Econ.* **42**(09), 113–126 (2023)
9. Zhu, R., Wang, D., Zhang, Y., et al.: Research and construction of intelligent control platform for new energy stations under digital transformation. *Technol. Ind.* **23**(15), 203–210 (2023)
10. Xu, Y.: Exploration and practice of national energy digital transformation. *China Standard.* **21**(08), 36 (2023)
11. Liu, J., Wang, L.: Analysis of influencing factors and countermeasures for digital transformation of energy enterprises under the background of dual carbon Economic. *Outlook of Bohai Rim* **22**(03), 13–15 (2023)
12. Wang, Y., Wang, J.: Experience, challenges, and suggestions for digital transformation of energy enterprises. *China Energy* **44**(11), 28–35 (2022)



Research Progress on the Infiltration of Digital Finance and Relative Deprivation of Consumption

Kenjie Chen¹, Shaohui Zhang²(✉), and Chuyuan Jiang¹

¹ Guangdong University of Science and Technology, Dongguan, China

² Guangdong University of Finance, Guangzhou, China

shaohuiemail@126.com

Abstract. Consuming power relative deprivation serves as a metric for measuring consumer inequality, providing a direct insight into the authentic disparities in welfare between residents. The effective operation of the financial system facilitates the individual's optimization of expected income distribution through lending, and it can also alleviate the consumption gap between different groups by enhancing the purchasing power of the less affluent. Digital finance, with its ability to shatter the constraints of time and space, low barriers to entry, seamless trading, and low transaction costs, can effectively expand financial inclusion and penetration, enhance the efficiency of financial resource allocation, thereby augmenting people's sense of happiness. The pervasive influence of digital finance on consumption is primarily grounded in two facets: research into the effects of digital finance on consumer expenditure and consumption structure, and the impact of household debt on both expenditure and structure. The article concludes with a summary of the progress in the study of digital finance's encroachment and the relative deprivation of consumption, also providing insights into future research issues.

Keywords: Digital Finance · Relative Deprivation in Consumption · Sense of Happiness · Common Prosperity · Kakwani Index

1 Introduction

In the report of the Party Congress in October 2017, General Secretary Xi Jinping pronounced this significant judgment: “The principal contradiction within our society has shifted to the conundrum between the growing needs of the nation for a better life and the imbalances and insufficiencies of development.” The shift from the traditional “needs of material culture” to the present “desire for a better life” signifies a profound transformation in the resident demand framework and hierarchies [1] (Zhu Ziwen & Xu Mengyu, 2019). The economics community's investigation into the elements influencing national happiness predominantly revolves around income, the caliber of government, economic expansion, and social equity (Alesiona et al., 2004 [2]; Ng, 2008 [3]; Chen, et al., 2012 [4]; Wu, 2017 [5]; Xuan, et al., 2016 [6]). The renowned “Easterlin Paradox” illuminates

the inescapable truth that affluence does not inevitably lead to an elevation in one's well-being [7] (Easterlin, 1974). A significant theoretical explanation is centered around the "comparative perspective," wherein the individual's subjective welfare is intimately tied to their relative standing. Clearly, the deceleration of individual consumption provides a concomitant relief to relative poverty, fosters social equity and justice, augments the well-being of daily life, and upholds societal stability. Yet, the concept of relative consumption deprivation refers not to an absolute measure of consumption but rather to a comparative one, whereby individuals are compared with others in their reference groups who possess higher levels of consumption. The resulting relative consumption standing or condition can, to some extent, serve as a metric for measuring individual-level inequality [8] (Fehr and Schmidt, 1999). Consumption is not only the immediate embodiment of residents' enhanced quality of life, but also a vital foundation for achieving economic high-quality development. Yet, the emotions generated by social comparison psychology significantly impact consumer behavior [9] (Jacob et al., 2020), with the possible root of this influence being the relative deprivation phenomenon among residents. In the milieu of China's new era of socialism with Chinese characteristics, the steadfast commitment to "the people at the center" and the emphasis on "the comprehensive development of humanity" are inevitable prerequisites in resolving the principal societal conflicts. Ultimately, "to perpetually enhance the full development of humanity and common prosperity for all" and "to strengthen the sense of fulfillment, contentment, and security among the people" remain the ultimate objectives. Consumption relative deprivation provides a metric that can, to a certain extent, reflect the authentic disparities in welfare among residents; it distinctly portrays the level and manifestations of residential happiness, allowing for a more nuanced understanding of the Chinese citizens' aspirations for daily well-being. The "eighty-two" pattern of Guangdong's economic development yet remains unsightly turned, as illustrated in Fig. 1. From 2019 to 2023, the combined GDP of the nine cities within the Guangdong-Hong Kong-Macao Greater Bay Area consistently exceeded the economic total of the other twelve prefecture-level cities within Guangdong. Within the span of these five years, the economic prowess of the Guangdong-Hong Kong-Macao Greater Bay Area has been robust, manifesting in a steady ascension in the aggregate GDP, whereas although the other twelve prefecture-level cities also exhibited growth in their GDP total, it did so at a slower pace, with the totals still vastly trailing behind. Please consider the following translation:

Behold the graph below Fig. 2, which illuminates the annual escalation in the economic prowess of nine cities within the Guangdong-Hong Kong-Macao Greater Bay Area, from 2019's 80.71% to 2023's 81.25%. This pattern signifies a robust upward trajectory. The economies of the other twelve provincial cities progressively diminish year by year, from 19.29% decline to 18.75%, a marked distance short of the ultimate objective of "collective prosperity for all". In 2023, General Secretary Xi Jinping emphasized during his tour of Guangdong that the essence of China-style modernization lies in the collective wealth of all citizens, while the inevitable prerequisite for achieving shared prosperity is the harmonious development of regions.

Guangdong, situated on the eastern coastline, stands at the forefront of China's economy, boasting the longest unbroken record in economic aggregate for 34 years and commanding nearly 9% of the nation's total populace of 127 million. The pervasive advent

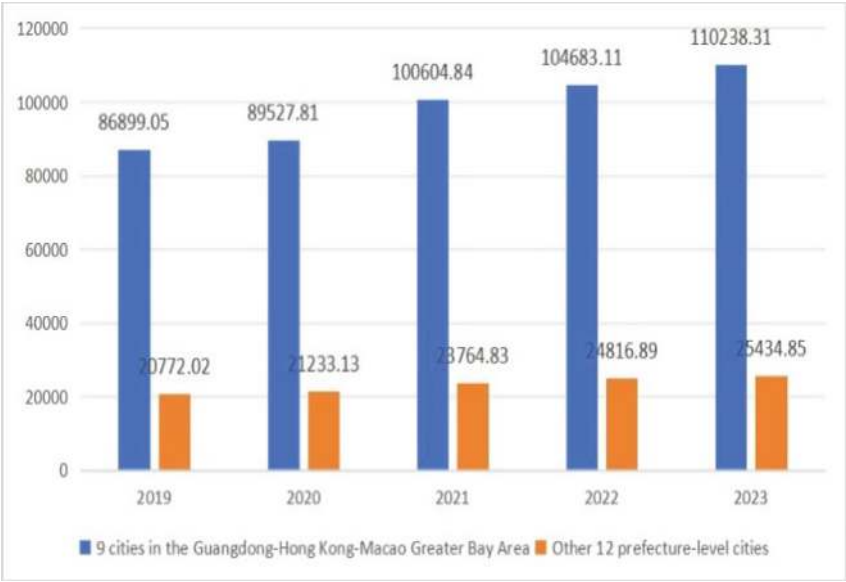


Fig. 1. 2019–2023: A comparison chart of the economic total of the nine cities in the Guangzhou-Hong Kong-Macao Greater Bay Area to that of twelve prefecture-level cities.

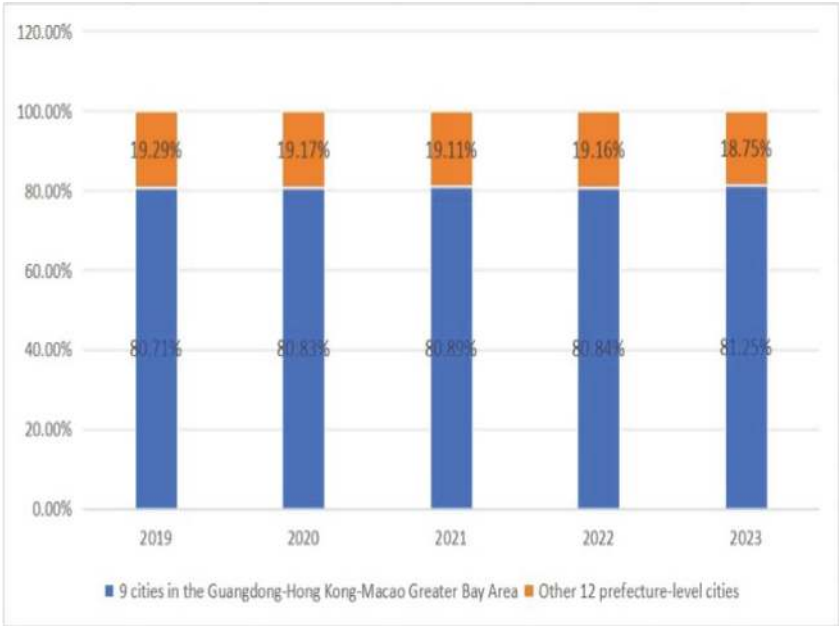


Fig. 2. Portrays a Contrast of the Economic Total of the 9 Cities in the Guangzhou-Hong Kong-Macao Greater Bay Area to that of the 12 s-Level Cities

of digital finance may exponentially augment residents’ sense of happiness; its potential to alleviate disparities between regions; and its capacity to aid in achieving President Xi’s ultimate vision of shared prosperity. These are pivotal issues affecting the welfare of 127 million people. Properly addressing these issues benefits not only the lives of these individuals but also that of the nation as a whole, assuming a significant demo effect across the nation. The “Digital Economy Development Blueprint for the 16th Five-Year Plan” promulgated by the National Development and Reform Commission highlights the rapid expansion of China’s digital economy, yet the issues of unbalanced development, insufficiency, and lack of standardization remain prominent. Urgent transformation is imperative to expedite addressing bottlenecks and vulnerabilities, augmenting the governance capabilities in the digital sphere, and charting a path of high-quality advancement unique to our nation. In July 2022, the inaugural release of “Guangdong’s Guides for Digital Economic Development 1.0” was formally unveiled. The document posits digital industrialization as the core tenet of developing digital economy, serving as a pivotal impetus in propelling economic high-quality advancement.

Presently, common measures for gauging the disparity of consumption lie in the commonly utilized Yitzhaki index, Kakwani indices, and Podder indices. The selection of the Kakwani index as a metric for measuring individual relative deprivation is premised on inherent imperfections within both the Yitzhaki index and the Podder index. Specifically, it entails assuming a sample of n individuals with total consumption values denoted by $x_1, x_2, x_3, \dots, x_{n-1}, x_n$, wherein each individual’s expenditure differs; these can be arranged in ascending order, with $\bar{\mu}_{xk}$ representing the average of all consumptions surpassing x_k , γ_{xk} denoting the proportion of individuals with expenses exceeding x_k , and $\bar{\mu}$ standing for the overall average. The formula for this indicator is as follows:

$$RD(x, x_k) = \frac{1}{n\mu_x} \sum_{i=k+1}^n (x_i - x_k) = \gamma_{xk}^+[(\mu_{xk}^+ - x_k)/\mu_x] \tag{1}$$

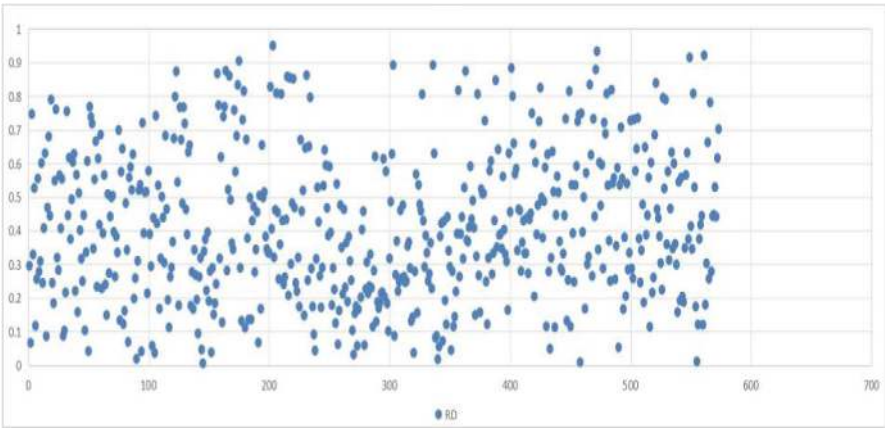


Fig. 3. Portrays the relative deprivation of residents in Guangdong to consumer spending in 2020.

Figure 3 delineates the scatter plot representing the residents of Guangdong Province' relative deprivation in consumer spending for the year 2020, with data sourced from the China Social Science Research Center of Peking University, which provided the data from the China Family Panel Studies (CFPS) on residential expenditure. The above-depicted image clearly illustrates that, at the 0.5-interval mark, scatter plots are densely situated beneath the 0.5 axis, whereas scatter plots on the axis above are comparatively scarce. This indicates that in 2020, the disparity in consumer expenditure among residents of Guangdong was of grave significance, with the necessity for enhancing the resident's sense of happiness.

2 Digital Finance's Influence on Consumer Expenditure and Consumption Structure

The influence of digital finance on consumption is primarily concentrated within the realm of research into how it impacts consumer expenditure and the composition of consume. Yi, Xingyuan et al. (2018) propose that digitally dispersed inclusive finance can enhance residential consumption by two mechanisms: the relief of liquidity constraints and the facilitation of payment convenience [10]. Digital finance exerts a siphoning effect on consumer expenditure [11] (Wang Haijun, Yang Hu, 2022); an efficacious operation of the financial system aids individuals in optimizing the distribution of expected income through lending; it also alleviates inter-group disparities in consumption by enhancing the purchasing power of low-income communities. Jiang Hongli and Jiang Pengcheng (2020) conducted empirical research using a dynamic panel model, uncovering that digital inclusive finance fosters an elevation in residential consumption levels and an upgrading of consumption structures, with a more significant impact on the advancement of luxury consumption than its counterpart. Furthermore, the income disparity between urban and rural areas and the industrial structure plays a mediating role within this process [12]. The investigation conducted by Cai in 2021 reveals that digital inclusive finance has a significant impact on per-capita consumer expenditure within urban and rural populations, with its influence on the per-capita consumption expenditure of rural residents outweighing that of urban residents [13]. Digital finance has profoundly altered the modes of residential consumption and consumer psyche, significantly diminishing the cost of shopping for consumers, enhancing the consumption experience, and giving birth to a succession of new demand for consumption [14] (Zhang et al., 2020). Moreover, digital finance boasts significant universality and inclusivity attributes, not only optimizing the allocation of financial resources but also propelling the decline of service and product thresholds in finance, thereby furnishing the vulnerable communities that have been overlooked by traditional banking with differentiated and multi-tiered financial services. The measure aims to ease the budgetary constraints faced by residents, thereby mitigating their fluidity constraints and facilitating the smoother intertemporal consumption of resident expenditure, releasing repressed consumer demands; it also alleviates financing constraints for small and medium-sized enterprises, propels the optimization and upgrading of industrial structures, and cultivates new forms of consumption oriented by residential demand, thus stimulating consumption potential [15] (Li Chunfeng & Xu Yaixuan, 2022). The existing research on the impact of digital finance on consumer

spending primarily analyzed two dimensions: one focused on the heterogeneity of digital finance's influence across various consumption categories, largely relying on macro-level data. However, due to the varying definitions of consumer expenditure categories, the concordance of research findings remains uncertain. The principal elements encompass: the digital finance's impact on the structure of consumption distinction; the influence of digital finance on regional disparities in consumption; and the discrepancies between urban and rural populations in the effects of digital finance on consumption. Explore the micro-dynamics of digital finance's impact on residential consumption, primarily utilizing micro-level data, with a conceptual framework drawn from the traditional consumer function theory, presenting a distinct perspective on the research. Emphasize: Digital finance facilitates payment, emblematic of the direct stimulation of consumption [14] (Zhang Xun et al., 2020).

3 The Influence of Family Debt on Consumption Expenses and Consumer Structure

There are studies investigating the impact of household debt on consumption expenses and consumer expenditure structures. The team of Ma Yuancong et al. (2020) pointed out that online lending has intensified the surge in household debt's leverage ratios and might diminish solvency for debt repayment [16]; a study conducted by Zhang Yalin and Yao Linzhen (2020) utilizing CFPS data also revealed that moderate indebtedness can effectively alleviate individual-level disparities in consumption inequality. Many academic works delve into the interconnections between household debt and consumption from a broad economic perspective [17]. The team led by Mian and Sufi (2010) uncovered that a surging rate of homeownership among the higher income brackets was a pivotal factor in the global financial crisis's impact on retail spending [18]. Krugman (2012) posits that the removal or diminution of leverage or the expectation of such a reduction may lead to a substantial decline in the consumer expenditure of debtors, with the augmentation in spending by savers or creditors not sufficient to counteract the magnitude of this decrease. Consequently, de-leveraging exacerbates the economic recession of a nation. Lombardi and his peers (2017) have demonstrated from an empirical standpoint that short-term household debt can bolster the economy, yet in the long run, family debt hinders economic expansion. Both Drehmann and Juselius (2012) and Juselius and Drewer-Müller (2015) have discerned the adverse impact of debt on credit expansion and expenditure growth, suggesting a propensity towards a future economic recession or a financial crisis. Another segment of the literature delves into how families' debt obligations impact their consumer behaviors from a micro-family perspective. Dynan and Edelberg (2013) utilize panel data regression analysis based on the consumer finance survey data from the United States in the years 2007–2009 to disclose that an increase in household debt levels corresponds with a decrease in automatic expense reduction by families due to repayment obligations, which stems from financing constraints within households and uncertainty surrounding the future economy [19]. Kim and Min (2016) employed micro-data from the Korean family across two epochs, spanning the years 2007–2009 and 2011–2013, to compare the conspicuous disparities between high leverage and low leverage households in their consumption expenditure. The study unveiled

that families of higher leverage experienced sluggish growth in their purchasing power due to the greater uncertainty in future credit access inherent in high-debt burden, compelling them to curtail expenses to diminish this uncertainty [20]. Baker (2018) posits that, owing to liquidity constraints, the consumption expenditure of high-debt-ratio families is more heavily impacted by debt shocks [21]. The Kovacs et al. (2018) employ data from the family unit to discern that families of high leverage are more profound in their expenditure curtailment during the financial crisis than those of average leverage [22]. Gong Yuna, Wei Nan, and Cai Tiantian (2021) utilizing data from the Chinese Household Survey in 2015 and 2017, conclusively demonstrate that household debt in China does not significantly augment or diminish total household expenses or consumer spending, yet it amplifies the role of family wealth in enhancing consumption. Household liabilities constraints stunt development and pleasure-seeking consumption; the inhibitory effect of household debt on the recessive expenses of consumer goods for low-income families is amplified, while the amplifying influence of wealth consumption on the reinforcement of such effects is more pronounced among high-income families [23].

4 Review and Expectations

The literature is either researching the influence of digital finance on consumer expenditure and consumption structure or addressing the impact from the perspective of household debt. Rare are scholars who have delved into the impact of digital finance on individual happiness. Qian Xuesong, Yuan Zengyong (2022) utilized the Chinese Digital Universal Access Finance Research Center at Peking University to craft the China Digital Universal Access Finance Development Index and data from the China General Social Survey (CGSS) to theoretically explore and empirically verify the influence of digital universal access finance on residents' subjective happiness from a micro perspective. The study reveals that the advancement of digital inclusive finance has profoundly enhanced the well-being of residents, with indices measuring the broadened scope and depth of use of digital inclusive finance, as well as the extent to which digital support services are available, all exude a pronounced positive impact upon residents' subjective happiness. The principal measure of residents' subjective happiness within this document is derived from the questions on personal fulfillment during the CGSS survey: "In general, how would you describe your life as happy?" Options include "Extremely unhappy," "Relatively unhappy," "Neither unhappy nor satisfied," "Relatively happy," and "Extremely happy," each assigned a score between 1 and 5, with higher scores denoting a greater intensity of residents' subjective happiness. This evaluation method, predominated by sentiment, yields a certain degree of randomness that can impact the objectivity of the research findings [24]. A solitary contingent of scholars in this nation has begun to utilize the objective quantified metric of consumer deprivation to investigate national happiness, yet research into the role of digital finance in mitigating this disparity is only now emerging, with a paucity of literature addressing this theme. Furthermore, there exists no literature that employs the tool variable of family leverage rate to examine the impact of digital finance on consumption disparities.

Acknowledgment. This dissertation is generously funded by the research topic “Digital Finance’s Percolation and the Relative Denial of Consumption—A Channel Analysis Based on Household Leverage” from the Guangdong Science and Technology University.

References

1. Zhu, Z., Xu, M.: Changes in China’s economic structure and high-quality development - summary of the First China development economics scholars forum. *Econ. Res.* (3) (2019)
2. Alesiona, A., Di Tella, R., Macculloch, R.: Inequality and happiness: are Europeans and Americans different? *Public Econ.* **88**(9–10), 2009–2042 (2004)
3. Ng, Y.: Happiness surveys: some comparability issues and an exploratory survey based on just perceivable increments. *Soc. Indic. Res.* **38**(1), 1–27 (2008). <https://doi.org/10.1007/BF00293784>
4. Gang, C., Shu, L.: How can the government make people happy? An empirical study on the impact of government quality on residents’ well-being. *Manag. World* (8), 55–67 (2012)
5. Wu, R.: Urbanization, provision of basic public services, and residents’ subjective well-being: an empirical analysis based on microdata from 56 cities. *Popul. Dev.* (3), (2017)
6. Ye Xuan, Yongze Yu: Public expenditure structure, public services and residents’ well-being. *Labor Economic Research* (5), (2016)
7. Easterlin, R.A.: Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz. Academic Press, New York (1974)
8. Fehr, E., Schmidt, K.M.: A theory of fairness, competition and cooperation. *Q. J. Econ.* **114**(3), 817–868 (1999)
9. Jacob, I., Khanna, M., Rai, K.A.: Attribution analysis of luxury brands: an investigation into consumer-brand congruence through conspicuous consumption. *J. Bus. Res.* **116**, 597–607 (2020)
10. Yi, X., Zhou, L.: Does the development of digital inclusive finance significantly affect household consumption? Micro evidence from Chinese households. *Finan. Res.* **11**, 47–67 (2018)
11. Wang, H., Yang, H.: The penetration of digital finance and the expansion of household debt in China - based on the transmission mechanism of mortgage and consumption. *J. Wuhan Univ.* (1) (2022)
12. Jiang, H., Jiang, P.: Research on the effect of digital inclusive finance on the improvement of residents’ consumption level and structural optimization. *Mod. Finan. Econ.* (J. Tianjin Univ. Finan. Econ.) **10**, 18–32 (2020)
13. Cai, P., Cai, G., Wang, W.: Does digital inclusive finance promote the consumption upgrade of urban and rural residents? An empirical analysis based on provincial panel data. *J. Hubei Univ. Econ. (Soc. Sci. Ed.)* **18**(07), 33–38 (2021)
14. Zhang, X., Yang, T., Wang, C., et al.: Digital financial development and household consumption growth: theory and China’s practice. *Manage. World* **11**, 48–63 (2020)
15. Li, C., Xu, Y.: Research on the effect of digital finance driving residents’ different structural consumption. *Mod. Finan.* (4) (2022)
16. Ma, Y., Hong, H., Lin, L.: Reverse incentives for borrower performance in online lending: empirical evidence from the data of “Renren Dai”. *Finan. Res.* (5) (2020)
17. Zhang, Y., Yao, L.: Family debt and relative deprivation in consumption: based on the perspective of housing debt and non-housing debt. *Finan. Res.* (8) (2020)
18. Mian, A., Rao, K., Sufi, A.: Household balance sheets, consumption, and the economic slump. <https://doi.org/10.2139/ssrn.1961211>. Accessed June 2013

19. Dynan, K., Edelberg, W.: The relationship between leverage and household spending behavior: evidence from The 2007–2009 survey of consumer finances. *Fed. Reserve Bank St. Louis Rev.* **95**(5), 425–448 (2013)
20. Kim, Y.L., Min, H.: Household debt and consumer spending in Korea: evidence from household data. *KDI J. Econ. Policy* **38**(4), 23–44 (2016)
21. Baker, S.R.: Debt and the response to household income shocks: Validation and application of linked financial account. *J. Polit. Econ.* **126**(4), 1504–1557 (2018)
22. Kavas, A., Rostom, M., Bunn, P.: Consumption response to aggregate shocks and the role of leverage. Working Paper (2018)
23. Gong, Y., Wei, N., Cai, T.: Household debt, consumption growth, and consumption upgrading of chinese residents: empirical evidence based on CHFS. *Contemp. Econ.* (7) (2021)
24. Qian, X., Yuan, Z.: Digital inclusive finance, residents' living pressure and well-being. *Econ. Horiz.* (1) (2022)



Analysis of the Impact of Information Technology on the Informatization of Management Accounting in the Digital Economy Era

Fanling Kong^{1(✉)} and Shuyan Meng²

¹ Guangdong University of Science and Technology of Finance and Economics,
Guangdong 523129, Dongguan, China
kongfunling@126.com

² Global Literary Industries Management, University of Southampton, Southampton SO17 1BJ,
Hampshire, UK

Abstract. This study analyzes the impact of XBRL (eXtensible Business Reporting Language) technology on the development of enterprise management accounting informatization using financial data from 2015 to 2021 as research samples. The results indicate that financial reports prepared using XBRL technology have improved the level of AR to a certain extent, effectively enhancing the quality of financial information. Additionally, it has also improved the work quality of management accountants and provided a basis for decision-making by the enterprise management, playing a crucial role in promoting the healthy development of enterprises.

Keywords: XBRL Technology · Management Accounting Informatization · Market Model · Financial Reports

1 Introduction

In the era of the digital economy, the value of data in social and economic life is increasingly highlighted, providing corresponding pathways for the restructuring of economic structures and changes in lifestyles [1]. Especially in today's rapidly developing market economy, the competitive pressure on enterprises is gradually increasing, forcing them to undergo transformation and innovation. Management accounting, as the main support for enhancing the economic management capabilities and economic efficiency of enterprises, also faces unprecedented challenges. Therefore, in order for enterprises to continuously enhance their competitive capabilities, they need to fully leverage the role of information technology, achieve effective integration of information technology and management accounting, and continuously improve the level of management accounting informatization. XBRL technology, as a product that integrates information technology and accounting technology, can standardize financial information, establish or realize a common platform for the provision, acquisition, transmission, exchange, analysis,

and comparison of accounting information, facilitating efficient and rapid use of this information-by-information users, thereby achieving counting once, sharing resources [2].

In this regard, this article analyzes the impact of XBRL technology on the informatization of management accounting, aiming to provide a reference for enhancing the level of informatization of management accounting in enterprises.

2 Overview of Management Accounting Informatization and XBRL Technology

2.1 Management Accounting Informatization

Management accounting informatization is the key driving force of the transformation of modern enterprise management. Relying on the computer Internet platform, it integrates cost management, investment management, budget management and other dimensions, and realizes the comprehensive acquisition, accurate calculation and intuitive display of enterprise data. In this process, data acquisition is the foundation. Through the construction of efficient data collection mechanism, to ensure the real-time, accurate and comprehensive internal and external information, to provide a solid data support for analysis and decision-making. Model calculation uses preset advanced algorithms and models to deeply mine and analyze massive data, automatically generate key reports such as cost control, investment decision, budget implementation and so on, and reveal the operation status and potential risks of enterprises. In the data display process, through charts, dashboard and other visual means, complex data is transformed into intuitive and easy to understand information, helping decision-makers to quickly grasp the core points and make wise judgments. At the same time, management accounting information also emphasizes the importance of data security and system integration, to ensure the security of data in transmission and storage, as well as the seamless connection with ERP, CRM and other enterprise information systems, to promote data sharing and process collaboration, and improve the overall operational efficiency. In addition, the establishment of the continuous optimization mechanism enables enterprises to constantly adjust and upgrade the information system according to the market changes and management needs, and maintain its advanced and practical [3].

2.2 XBRL Technology

XBRL is a technology for exchanging and processing financial reporting information, which enables the preparation, disclosure, and utilization of corporate financial statements under established standards. It is a financial processing technology that combines traditional financial processing methods with modern data processing technologies. The XBRL technology architecture mainly includes technical specifications, classification standards, and instance documents. Technical specifications refer to the outline of XBRL technology, which includes corresponding working mechanisms and syntax rules and serves as the basis for establishing classification standards. Classification standards are an important part of technical implementation, consisting of patterns for defining relationships between relevant concepts and specifying how instance documents should be

created [4]. In the process of standard formulation, technical specifications should be used as the basis, with practical applications as the guide; instance documents need to be prepared in accordance with the requirements of technical specifications and classification standards. They serve as important carriers of XBRL data, presenting the data content externally during the application of XBRL. The relationship among the three components is illustrated in Fig. 1.

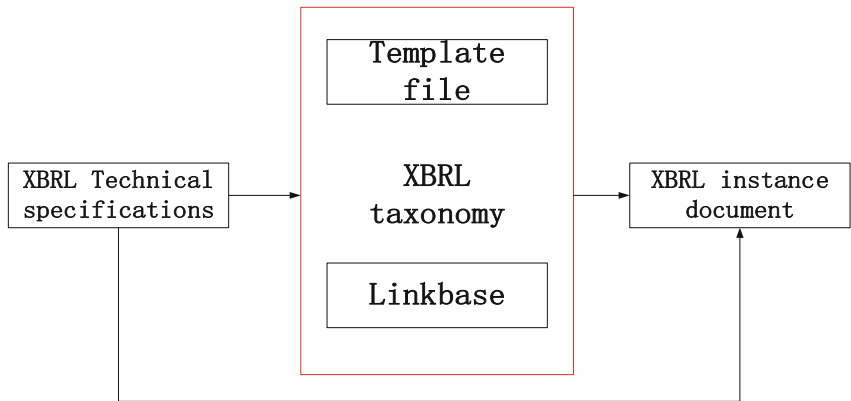


Fig. 1. XBRL Technology Architecture Relationship Model

3 Impact of XBRL Technology on Management Accounting Informatization

3.1 Case Selection

In this study, the financial data reported by a certain company from 2015 to 2021 was taken as the research sample to analyze the impact of XBRL technology on the management accounting informatization of the company. During the operation of this company, XBRL technology has been introduced in financial management. In compliance with national requirements and considering its own development, the company has completed the preparation and submission of XBRL network financial reports, making it a representative enterprise. Therefore, the research results are considered to be reasonably credible.

3.2 Research Hypotheses

In the management accounting information system, it can provide financial staff with stock prices presented by the stock exchange. Since stock prices reflect real-time changes in the stock market and the actual development of the company, changes in key indicators such as cumulative abnormal returns can be analyzed to assess the quality of financial information. This study mainly explores the impact of XBRL technology on management

accounting informatization. The crucial basis for management accountants in making analytical decisions is the financial information of the target company. Therefore, it is hypothesized that XBRL technology can effectively enhance the quality of financial information and also improve the efficiency of management accounting work [5].

3.3 Research Method

In this study, event study methodology is primarily employed. A specific event is selected, relevant data is collected, and the changes in stock returns before and after the event are calculated. Subsequently, the analysis is conducted on how this specific event influences the changes in stock prices and returns of the sample. This method is used to examine the extent to which the occurrence of a specific event affects price changes and the quality of information disclosure. Price fluctuations are often caused by changes in certain indicators in a company's financial reports [6]. XBRL technology can correspond financial indicators by setting classification standards, providing relevant corporate information to users, which may lead to corresponding stock price fluctuations. Therefore, the selected method in this study is considered feasible [7].

The event date is primarily the time point when the market accepts event information. As the hypothesis to be verified in this study is that XBRL technology can effectively enhance the quality of financial information, the date when the company reported the XBRL network financial report is selected as the event date. Additionally, since XBRL technology is an advanced digital information technology, network financial reports prepared using this technology need a certain period to fully function. If the window period is relatively long, stock price changes will be influenced by different factors, and if the window period is relatively short, the conclusions may lack persuasiveness [8]. Therefore, considering the actual development of the company and the hypothesis proposed in this study, the window period is set as $[-1, 1]$ for the subsequent research to ensure the conclusions drawn are persuasive.

3.4 Model Design

To analyze whether the disclosure of financial reports prepared using XBRL technology has significantly contributed to excess returns on the stock price of the company, it is necessary to calculate the expected returns on the stock price. In this research, the main methodology employed is the market price model, which is based on the correlation between market indexes and historical stock prices of the company [9]. By considering the changes in the index, the analysis aims to predict the future changes in the company's stock price. The specific steps are as follows:

Calculate the correlation index, such as the BETA value, between the historical stock prices of the company and a specific market index or industry index.

(2) Based on the calculation in (1), compute the daily stock returns of the company.

(3) Using the results from (2), calculate the fair value of the company's stock on a given day. The constructed market model can be represented as follows:

$$\begin{cases} R_{it} = \alpha + \beta R_{mt} + \varepsilon \\ \hat{R}_{it} = \hat{\alpha} + \hat{\beta} R_{mt} + \varepsilon \end{cases} \quad (1)$$

In Eq. (1), R_{it} represents the actual return of asset i during the estimation period t ; R_{mt} represents the actual return of a portfolio that includes all market securities in the estimation period t . By estimating α , β through linear regression, one can obtain the corresponding expected returns.

Based on the constructed market model, the cumulative abnormal return (CAR) is calculated using the event study method. The specific process is as follows:

Determination of dates. In this research process, the impact of XBRL technology on information disclosure quality needs to be examined. Therefore, the event date is set as the date of XBRL report disclosure, with window periods of $[-1, 1]$ and estimation periods of $[-30, -1]$.

(2) Data sources. To verify whether XBRL financial reports significantly impact the stock prices of listed companies, this study focuses on the disclosure of XBRL financial reports by a certain company from 2015 to 2021. The analysis of financial report information quality is based on data mainly sourced from the company's annual financial reports and the WIND database.

(3) Calculation of daily stock returns and market value returns. Based on the closing prices of stocks and market indices, daily stock returns and market value returns are calculated using the following formulas:

$$\begin{cases} R_{it} = (P_{it} - P_{i,t-1})/P_{i,t-1} \\ R_{mt} = (P_{mt} - P_{m,t-1})/P_{m,t-1} \end{cases} \quad (2)$$

In Eq. (2), P_{it} represents the closing price of stock i on day t ; P_{mt} represents the composite index of the market on day t .

Calculate the estimated coefficients. The formula for calculating the estimated coefficients is:

$$R_{it} = \alpha + \beta R_{mt} + \varepsilon \quad (3)$$

(5) During the window period, calculate the normal stock return $\widehat{NR}$, which can be represented by the following formula:

$$\widehat{NR} = \hat{\alpha} + R_{mt} * \hat{\beta} + \varepsilon \quad (4)$$

(6) Estimate the abnormal return rate $\widehat{AR}$ and the cumulative excess return rate $\widehat{CAR}$, the calculation formula can be expressed as:

$$\widehat{AR} = R - \widehat{NR} \quad (5)$$

$$\widehat{CAR} = \sum \widehat{NR} \quad (6)$$

4 Analysis and Discussion

4.1 Market Model Estimation

Through the analysis above, it is known that in this study, the calculation of stock price normal returns is mainly conducted through market models. The magnitude of returns is reflected through the comprehensive yield, and the first 30 trading days before the

disclosure of XBRL financial reports for each year are taken as the estimation interval to calculate the overall returns [10]. The OLS (Ordinary Least Squares) estimation of the market adjustment model for each year is conducted through Formula (3), with detailed results shown in Table 1.

Table 1. Estimation Results of Market Adjustment Model

Model		Dependent variable R_{it}				
		Unstandardized coefficient		Standardized coefficient		Significance
		B	Standard error	$Beta$	t	
2015	(constant)	0.060	0.150	——	0.390	0.700
	R_{mt}	0.540	0.160	0.530	3.320	0.002
2016年	(constant)	0.219	0.470	——	0.470	0.650
	R_{mt}	1.220	0.430	0.410	2.820	0.010
2017	(constant)	0.219	0.470	——	0.470	0.650
	R_{mt}	1.220	0.430	0.470	2.820	0.010
2018	(constant)	-0.140	0.260	——	-0.550	0.590
	R_{mt}	1.170	0.210	0.730	5.580	0.000
2019	(constant)	-0.140	0.230	——	-0.620	0.540
	R_{mt}	1.200	0.120	0.880	9.960	0.000
2020	(constant)	-0.089	0.179	——	-0.500	0.620
	R_{mt}	0.895	0.340	0.440	2.620	0.020
2021	(constant)	0.240	0.270	——	0.870	0.389
	R_{mt}	1.190	0.210	0.730	5.620	0.000

Based on the analysis of Table 1, it is found that the OLS estimate of R_{mt} is significant at the 5% level of significance, indicating a highly significant correlation between the individual stock returns of the company and the overall market returns. This suggests that the model chosen for this study is reasonable. Furthermore, the annual individual stock returns show a positive correlation with the overall market returns, indicating that individual stock changes are consistent with the market trends. However, there exists a certain difference in the reflection of individual stock returns relative to market returns variability [11].

4.2 CAR Estimation

After estimating the market model for each year, the returns of the stock before and after disclosing XBRL financial reports can be calculated using the model. The specific process is as follows: first, estimate the normal returns of individual stocks based on the constructed market model; second, calculate the estimated abnormal returns; Finally, by

overlaying $\widehat{NR}$ within the specified window period, one can obtain the CAR estimate [12].

At this point, the size of CAR within the interval $[-1,1]$ before and after the disclosure of XBRL financial reports by the company is shown in Fig. 2.

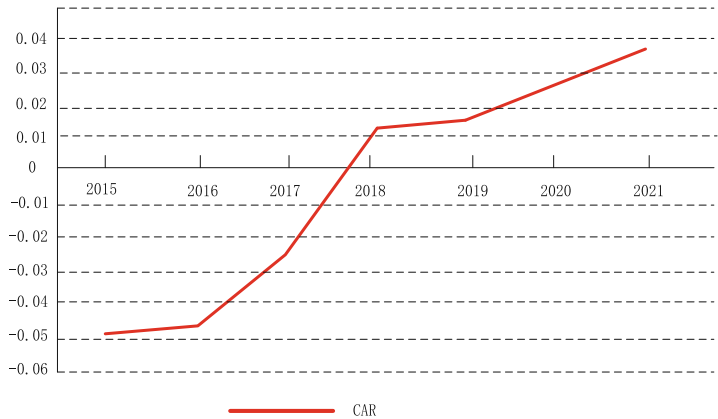


Fig. 2. Numerical Results of CAR

The financial reports of the company from 2015 to 2021, as shown in Fig. 1, clearly indicate a gradually increasing trend in CAR, suggesting that in the early stage of the company’s development, the importance of XBRL technology was not emphasized. Although the company established a management accounting information system during its development, it did not focus on the application of information technology, leading to the limited effectiveness of XBRL financial reports [13]. However, after the company applied XBRL technology to the management accounting information system, it has to some extent improved the quality of financial information, providing a basis for management accountants to analyze and make decisions [14].

4.3 Calculation of Excess Return Rate (AR)

The statistical characteristics of AR before and after the disclosure of XBRL financial reports by the company are detailed in Table 2.

Table 2. Statistical Characteristics of AR Before and After the Disclosure of XBRL Financial Reports (%)

Project	Before XBRL Financial Report Disclosure		After XBRL Financial Report Disclosure	
	Mean Value	Standard Deviation	Mean Value	Standard Deviation
Enterprise	−0.08690	0.978730	0.121280	1.731080

By analyzing Table 2, it is found that the average value of accounts receivable (AR) was negative before the disclosure of XBRL financial reports, and became positive after the disclosure, with a relatively large increase. This to some extent enhanced the abnormal returns of individual stocks [15]. Therefore, it can be concluded that the disclosure of XBRL financial reports effectively improved the accounts receivable of the company.

5 Conclusion

In this study, based on the empirical analysis framework of event study method, the sizes and changes of CAR and AR before and after the disclosure of XBRL financial reports by a certain company from 2015 to 2021 were analyzed. It was found that CAR showed a gradually increasing trend, and the mean of AR also exhibited a gradually increasing trend before and after disclosure, indicating that XBRL financial reports have to some extent enhanced the level of AR, demonstrating that XBRL financial reports can effectively improve the quality of financial information. This also indirectly confirms that XBRL technology can enhance the work quality of management accountants, provide a basis for decision-making by the management team, and play a very important role in promoting the healthy development of enterprises. Although this study confirms the impact of XBRL technology on the informatization of management accounting to a certain extent, due to the limitations of the researcher's level, it is difficult to comprehensively verify the impact of XBRL technology on the informatization of management accounting. There is still room for improvement in the article. Therefore, in future research, it is necessary to continue to focus on the advantages of XBRL technology and further verify the important role of XBRL technology in promoting the improvement of the level of management accounting informatization.

Acknowledgement. This research was supported by Foundation: Ph. D. Research Start-up Fund of Guangdong University of Science & Technology (Grant No. GKY-2024BSQDW-1).

References

1. Liu, H.: Research on the construction of management accounting informatization based on blockchain technology. *Contemp. Account.* **14**, 88–89 (2021)
2. Cormier, D., Dufour, D., Luu, P., et al.: The relevance of XBRL voluntary disclosure for stock market valuation: the role of corporate governance. *Can. J. Adm. Sci./Revue Canadienne des Sciences de l'Administration* **36**(3) (2018)
3. Kim, J.-B., Kim, J.W., Lim, J.-H.: Does XBRL adoption constrain earnings management? Early evidence from mandated U.S. filers. *Contemp. Account. Res.* **36**(4) (2019)
4. Huang, X.: The impact of information technology on the informatization of management accounting in the digital economy era. *Bohai Econ. Outlook* **5**, 25–27 (2023)
5. Mengqian, G.: Development path of management accounting informatization technology for public utility units in the big data era. *China Town Township Enterpr. Account.* **1**, 226–227 (2020)
6. Bhattacharya, N., Cho, Y.J., Kim, J.B.: Leveling the playing field between large and small institutions: evidence from the SEC's XBRL mandate. *Soc. Sci. Electron. Pub.* (2018)

7. Shanshan, H.: Development measures of management accounting informatization in the software information technology industry. *Bus. Story* **35**, 61–63 (2021)
8. Yang, Y.: The impact of information technology on the informatization of management accounting in the digital economy era. *Greater Wuhan* **24**, 130–132 (2022)
9. Rong, L.: Construction and application of management accounting system based on neural networks. *Bus. Account.* **12**, 83–87 (2023)
10. Emmanuel, G., Williams, A.A., Ivy, E., et al.: An analysis of the effects of management accounting services on the financial performance of SME: The moderating role of information technology. *Cogent Bus. Manage.* **10**(1) (2023)
11. Jinghui, M.: Problems and countermeasures in the development of management accounting informatization in China. *Finance News* **4**, 1–3 (2023)
12. Zhou, Y., Zhang, H.: Research analysis on the transformation of enterprise financial accounting to management accounting under the background of big data. *North South Bridge* **13**, 91–93 (2023)
13. Changyou, C.: The impact of accounting informatization on enterprise financial management and countermeasures. *Chin. Agric. Account.* **33**(2), 12–14 (2023)
14. Wang, J.: Analysis of the development path of financial management and accounting informatization in public institutions. *Brand Res.* **25**, 215–217 (2023)
15. Malik, A.A., Minh, N.N.: Impact of risk-taking tendency and transformational leadership style on the use of management accounting system information: a direct-mediation model. *Balt. J. Manage.* **18**(3), 366–385 (2023)



Research on the Information System Construction of Enterprise Human Resource Management Based on Integrity Culture

Weiling Hou^(✉) and Lili Xie

Shandong Business Institute, Yantai, Shandong, China
360676939@qq.com

Abstract. Human resource is one of the four resources of enterprises. Excellent human resource management can enhance the internal cohesion and centripetal force of enterprises, reduce the mobility of employees and internal friction of enterprises, and help enterprises survive and develop in the fierce market competition. The construction of honesty culture is the key to the service and management of human resources. This paper will analyse how to construct the integrity system in enterprise human resources service. In This paper, honesty culture is integrated into the construction of enterprise human resource management information system, including organizational management, recruitment management, performance management, salary management, training management and staff management. Through the actual operation results, it shows that the new system is significantly improved compared with the traditional system in terms of stability, user satisfaction, system ease of use, fairness perception, integrity training participation, performance evaluation fairness, query response time and the incidence of dishonest behaviour.

Keywords: Information System · Human Resource Management · Integrity Culture

1 Introduction

Integrity has become a necessary condition for the development of enterprises. Honesty is the basis for the survival of a person and even an enterprise, and it marks the quality of an enterprise. Honesty has become the most basic moral principle of enterprises. Integrity is an integral part of the core competitiveness of an enterprise, the cornerstone of the long-term development of a company, and an important manifestation of corporate culture, which should become an organic part of the long-term development strategy of an enterprise. If we give up honesty, enterprises may get some short-term benefits, but it will also seriously hinder the progress and development of enterprises. The same is true in the human resource management of enterprises [1]. The establishment of honesty and credit system in human resource management of enterprises will play a vital role in the success of enterprises. Although honesty is very important and prominent in human resource management, enterprises and employees in our country have shown dishonesty

to varying degrees, and even can be said that enterprises in our country are facing a crisis of honesty. In the final analysis, the problem of honesty and credit faced by enterprises is the problem of honesty and credit of people. To solve the problem of honesty and credit of enterprises, we should start with human resources management. At present, the human resource management information system of many enterprises has not fully integrated the elements of integrity culture, which to some extent limits the potential of enterprise human resource management. The purpose of this study is to explore how to integrate the integrity culture into the construction of enterprise human resource management information system, in order to improve the efficiency of the system and the management level of enterprises. Through the analysis of the existing practice of human resource management, this paper puts forward a set of human resource information system integrated with integrity culture, aiming at realizing the informatization, intellectualization and integrity of human resource management.

2 Design Requirements

2.1 Functional Requirements

In the current enterprise human resource management information system, the realization of functional requirements often ignores the integration of integrity culture. This not only affects the comprehensiveness and depth of the system, but also weakens the transparency and trust of enterprises in human resources management. The honesty problems of employing enterprises in the process of signing contracts mainly occur in the links of releasing demand information, investigating and signing contracts. Specifically, it can be manifested as dishonesty in information release, dishonesty in welfare commitment, dishonesty in job description, dishonesty in enterprise prospect description, etc. Job seekers' integrity problems in the process of signing contracts mainly occur in the links of publishing job information, accepting investigation and signing contracts. It is manifested by false certificates, false resumes and false certificates, which elevate themselves and belittle others. The realization of the functional requirements of the human resource management system integrated with the integrity culture should include the integrity culture training function, the company information sharing function, and the transparent reward and punishment mechanism. The system shall have integrity education and training modules to enhance employees' awareness and understanding of the importance of integrity through regular online courses and interactive activities; the system shall integrate integrity monitoring and evaluation mechanisms to detect and correct dishonest behaviour in a timely manner by recording and analysing employee's work performance and behaviour [2]. The system should provide a transparent information sharing platform to ensure that employees have access to company policies, rules and regulations, and work processes, thereby reducing misunderstandings and deception. Through the integration of integrity culture, human resource management system can not only improve the management efficiency of enterprises, but also establish a more solid foundation of trust between enterprises and employees.

2.2 Performance Requirements

Performance requirements focus on the stability, response speed and processing ability of enterprise human resource management information system in actual operation. In terms of data processing capability, an efficient human resource management system should be able to respond quickly to user requests, handle a large number of data input and queries, and ensure the accuracy and security of data. It must not be because of the integration of integrity culture, resulting in excessive system load, difficult to deal with a large number of user needs. The system can efficiently complete complex data analysis and report generation, and provide decision support for enterprises. In terms of ease of use, the human resource management system integrated with integrity culture should have a good user interaction experience, friendly interface, simple and intuitive operation, to ensure that users can quickly start and effectively use the system. In addition, the system should also have a high degree of security, through encryption technology and access control, to protect the data of enterprises and employees from unauthorized access or tampering [3]. Of course, information systems should be designed with redundancy mechanisms such as backup servers, failover, and data backup. This ensures the stability of the information system. By meeting these performance requirements, the system will be able to provide a stable, reliable and efficient human resource management platform for enterprises, while meeting the integrity culture, serving enterprises and employees, and helping enterprises gain an advantage in competition.

3 System Architecture

As shown in Fig. 1, the overall framework of the human resource management information system includes four layers: access layer, application layer, data layer and network layer. The system can be connected to the enterprise single sign-on platform, and can also support separate access to the system login entrance login system. The application layer is the core business system layer of enterprise human resource management, the access layer interacts with the application layer, and the application layer saves data to the system database through the data layer. The whole system runs in the enterprise intranet, which ensures the normal office needs of employees and increases the security of the system.

Compared with the ordinary human resource management system, we have integrated the integrity culture into the organizational management, recruitment management, performance management, salary management, training management, staff management and other modules. In the organizational management, by ensuring the transparency and fairness of the organizational structure, employees can clearly understand the company's hierarchy and job requirements; in the recruitment management module, the integration of integrity culture is reflected in the fair and just recruitment process [4]. In the performance management module, an objective and fair performance evaluation system is integrated into the integrity culture to encourage employees to honestly reflect their work achievements and problems. In the salary management module, we establish a salary system based on performance and contribution to show the integrity culture of the enterprise. In the integrity management module, training courses related to integrity are provided to educate employees to practice integrity in their daily work. In the employee

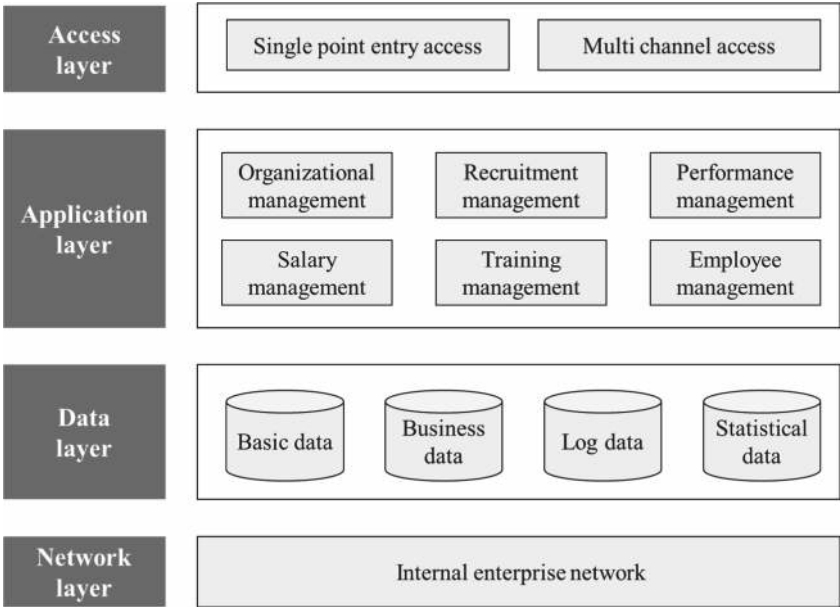


Fig. 1. Architecture of human resource management information integrated integrity

management module, employees are helped to realize the value of integrity in personal development by recording their work behaviour and performance.

4 Module Design

4.1 Organization Management Module

Human resource management information system can dispatch personnel internally to set up new departments, automatically evaluate positions and ranks after personnel dispatch, and quickly build an organizational system. In the change of department, when the department is added or merged, it is necessary to dispatch the original personnel or recruit new employees [5]. In the Organizational Management module, the responsibilities and expectations of each position should be clearly defined and documented. This helps create a transparent and predictable work environment where employees understand their roles and responsibilities. Ensure that all employees are fairly placed in appropriate positions based on their abilities, experience and performance through an honest position allocation mechanism. This helps prevent cronyism and unfair promotion and strengthens employees' trust in the organization [6]. The organizational management module should provide a clear career development path, so that employees can understand how to achieve job promotion through hard work and achievement. This transparency helps motivate employees and reduces the mistrust that comes from uncertainty. The module contains an integrity-based employee evaluation system that ensures fairness and consistency in the evaluation process. This includes an objective evaluation of the employee's performance, capabilities, and potential. Employees are encouraged to provide feedback

and suggestions through the organizational management module, establishing an open and honest communication channel. This helps management understand what employees really think and make improvements accordingly.

4.2 Recruitment Management Module

The whole process of recruitment demand initiation, resume acquisition, interview process, offer issuance and personnel entry can be completed online to achieve paperless operation [7]. Obtain the talents needed for the development of enterprises, introduce comprehensive personnel with high comprehensive quality and strong plasticity, and provide guarantee for the further training and selection of internal talents. Establish a clear recruitment policy that stresses the importance of integrity and ensures that all recruitment activities adhere to this principle. This includes treating all candidates fairly, without discrimination based on race, gender, age or other non-job-related factors. Ensure that every step of the recruitment process is open and transparent, allowing candidates to understand where they are in the recruitment process and what to expect next. This transparency helps to build trust in the candidate's business. Provide accurate and thorough information in the job description, including job responsibilities, requirements, and expected working conditions. Avoid exaggerated or misleading information and ensure that the candidate has a true perception of the position. During the hiring process, a thorough background check is performed, including verification of educational background, work history, and professional qualifications. This can help identify dishonest candidates and ensure recruitment of people who truly meet the requirements of the position. During the interview, the interviewer should demonstrate integrity and professionalism by providing true and accurate information about the company and the position. At the same time, candidates are encouraged to express their views and experiences honestly and avoid exaggerating or misleading. Provide a feedback channel for candidates to provide feedback on any questions or concerns during the hiring process. This helps companies to understand and improve the recruitment process and improve integrity.

4.3 Performance Management Module

The current enterprise performance appraisal work needs to be strengthened, through the performance appraisal can stimulate the enthusiasm of employees, through the quantification of work efficiency, can let employees do a good job of their own positioning, encourage employees to constantly learn to improve their ability, think about innovation in the work to improve efficiency, performance and salary associated with the current backbone of the enterprise. Promote the healthy and virtuous circle of production and operation activities of enterprises, and ensure the realization of business objectives of enterprises. Through a reasonable performance appraisal system, the performance of employees can be identified and evaluated fairly and comprehensively, which can identify the contribution of employees, stimulate their innovative spirit, improve their working methods, develop their potential abilities, enhance their integrity to the enterprise, and realize the double harvest of enterprise economic benefits and personal benefits [8]. Give honest, correct and effective feedback on the evaluation results. Through feedback,

on the one hand, both sides of the evaluation can reach an agreement on the performance of the examinee, on the other hand, employees can have a clear understanding of their strengths and weaknesses, which is not only conducive to the formulation of reasonable performance improvement plans by the examinee, but also can stimulate their ambition and work enthusiasm, so as to improve the overall performance of the enterprise. We will fully apply the evaluation results to the actual work to further improve and enhance employee performance. Performance appraisal is not an end in itself, but a means. Attention should be paid to the application of evaluation results, and at the same time, the effectiveness of performance evaluation can be tested.

4.4 Salary Management Module

Salary management is an important part of personnel management. In the daily work, salary management takes up a lot of time of personnel specialists, and the work has the characteristics of strong periodicity and repeatability, mainly including two aspects of salary and welfare management. The salary management module can help the salary administrator to realize the information management of the whole process from the formulation of salary strategy to the payment of salary, realize the efficient, convenient and reliable management of salary, and reduce the duplication of labour of personnel specialists. When a person anticipates that an action will bring about a desired result for the individual [9]. And when the result is attractive to the individual, the individual takes this particular action. The same is true of the employee's expectations of pay. However, if there is a big gap between his gains and his expectations, his job satisfaction and loyalty to employers will be reduced, which will lead to a decline in labour productivity and affect the development of enterprises. When an employee receives an income, he will compare the ratio of his income to his contribution with the ratio of other people's income to their contribution: if the two ratios are the same or similar, he will feel that it is fairer; However, if the former is less than the latter, it will produce a sense of unfairness, leading to a decline in the integrity of the enterprise, thus affecting their enthusiasm for work. It can be seen that the salary level, salary structure and salary form of enterprises will directly affect the integrity of employees to enterprises.

4.5 Training Management Module

Business owners should take the initiative to accept integrity training, mainly to improve their integrity, set an example, take the lead in setting an example, so that they can become the integrity model of employees, on this basis, shape corporate culture, corporate spirit and corporate ethics, and turn training management into the main position of building corporate culture. Develop specialized integrity training courses covering corporate values, ethics, compliance, and case studies of honest behaviour. These courses should be designed to be interactive and participatory to increase employee engagement and learning. Ensure that all employees are trained in integrity, regardless of their position or level [10]. Make a culture of integrity a part of every employee's career through online or face-to-face training. During the training process, use examples or cases of integrity within the enterprise to demonstrate the importance of honest behaviour. This helps employees understand the application of integrity in practical work. At the end

of the training, employees' ability to understand and apply the concept of integrity is assessed by means of tests, questionnaires or discussions. At the same time, collect feedback from employees on training content and methods to continuously improve the training plan. We incorporate integrity education into our employees' ongoing development programs, regularly updating and providing new integrity training materials to address changing work environments and challenges. At the same time, employees are encouraged to apply the knowledge learned in honesty training in practical work, and to practice honesty decision-making in a safe environment through simulation scenarios or role-playing. We reinforce the positive impact of a culture of integrity by rewarding and recognizing employees who demonstrate honest behaviour after training.

4.6 Staff Management Module

This module can complete the whole personnel cycle management of employee entry, transfer and transfer, form a digital employee file database, meet the analysis needs of enterprise managers and decision makers, present employee files through personalized business data, and comprehensively realize the data standardized management of enterprise personnel files. In the employee management module, the personal information, work experience and educational background of employees are recorded to ensure the authenticity and accuracy of all information. This helps to build up the integrity profile of employees and provide reliable data support for future evaluation and development. In the process of performance evaluation, objective and fair standards are adopted to ensure the accuracy and consistency of the evaluation results. Encourage employees to be honest about their work results and problems, and managers should also be honest about their performance. Provide employees with clear career development paths and promotion opportunities to ensure the fairness and transparency of the career development process. Through honest career planning and guidance, help employees achieve the consistency of personal goals and corporate goals. Establish an honest communication channel that encourages employee input and suggestions while ensuring that management is responsive to employee concerns. This helps build trust between employees and management. In the employee management module, the reward for honest behaviour and the punishment for dishonest behaviour are clearly defined. Encourage employees to abide by corporate values by rewarding honest behaviour; By punishing dishonest behaviour, we can maintain the integrity standards of enterprises.

5 Operation Results

We will demonstrate the positive changes and significant effects brought by the enterprise human resource management information system integrated with integrity culture compared with the traditional system through the actual operation results. This chapter will make a comparative analysis from multiple dimensions, and intuitively show the differences between the two systems in practical application through data tables. The comparative data are shown in Table 1.

The design of the new system pays attention to the response speed and data processing ability of the system, which ensures the high efficiency of the system operation while

Table 1. Comparison of data between the old and new systems.

Number	Indicators	Traditional system	New system	Percent improvement
1	System stability	0.96	0.99	2.4%
2	Customer satisfaction (5-point scale)	3.50	4.70	34.3%
3	System ease of use (5-point scale)	3.20	4.50	40.6%
4	Perception of fairness (5-point scale)	3.00	4.60	53.3%
5	Integrity Training Participation	0.64	0.91	42.2%
6	Fairness of performance evaluation (5-point scale)	3.40	4.70	38.2%
7	Query Response Time	5.20	3.80	26.9%
8	Incidence of dishonest behaviour	0.05	0.02	70.6%
9	Staff turnover rate	0.12	0.08	33.3%

integrating the elements of integrity culture. By optimizing the algorithm and database structure, the new system shows faster query speed and higher stability when dealing with large amounts of data. The new system has significantly improved user satisfaction by providing a more transparent and fair management process. Employees highly rated the ease of use and fairness of the system, which was reflected in the employee questionnaire. The implementation of the new system has had a profound impact on the construction of honesty culture in enterprises. Through regular honesty training and supervision mechanism, employee's sense of honesty has been strengthened, and the incidence of dishonest behaviour has been significantly reduced. The optimization of performance and salary management in the new system makes the performance evaluation of employees fairer, the salary distribution more reasonable, effectively reduces the turnover rate of employees, and improves the job satisfaction and loyalty of employees. The new system has a significant positive impact on the construction of corporate culture, the promotion of employer's integrity awareness and the management of performance and compensation. These changes directly promote the overall performance and sustainable development of enterprises.

6 Conclusions

This paper constructs and implements a set of enterprise human resource management information system with integrity culture. The new system shows significant improvements in several key indicators, including system stability, user satisfaction, system ease of use, perception of fairness, participation in integrity training, fairness of performance evaluation, query response time, and incidence of dishonest behaviour. These improvements not only improve the job satisfaction and loyalty of employees, but also effectively reduce the turnover rate of employees, and promote the overall performance and sustainable development of enterprises. The implementation of the new system has a significant positive impact on the construction of corporate culture, the improvement of employee's integrity awareness and the management of performance and salary. The results show that the integration of integrity culture is very important to improve the quality and efficiency of human resource management, which provides valuable reference for other enterprises. We will continue to optimize the functions of the system and strengthen the construction of integrity culture in order to achieve long-term and stable development of enterprises.

Acknowledgements. Integration of integrity culture into the entire process of enterprise operation plan design consulting (2024HXZX039).

References

1. Vrontis, D., Christofi, M., Pereira, V., et al.: Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *Int. J. Hum. Res. Manage.* **33**(6), 1237–1266 (2022)
2. Strohmeier, S.: Digital human resource management: a conceptual clarification. *German J. Hum. Resour. Manage.* **34**(3), 345–365 (2020)
3. Chowdhury, S., Dey, P., Joel-Edgar, S., et al.: Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Hum. Resour. Manage. Rev.* **33**(1), 100899 (2023)
4. Collins, C.J.: Expanding the resource based view model of strategic human resource management. *Int. J. Hum. Resour. Manage.* **32**(2), 331–358 (2021)
5. Manuhutu, M.A., Uktolseja, L.J., Sitaniapessy, S.F., et al.: Implementation of the general administrative management information system at victory university of Sorong. *IJISTECH (Int. J. Inf. Syst. Technol.)* **5**(2), 136–142 (2021)
6. Albarqi, A.H., Yuniarthe, Y., Hendri, R.: Employee workload data processing information system using Android-based time motion study method. In: *Proceedings vol. 1, no. 2*, pp. 30–35 (2022)
7. Malik, S.Y., Cao, Y., Mughal, Y.H., et al.: Pathways towards sustainability in organizations: empirical evidence on the role of green human resource management practices and green intellectual capital. *Sustainability* **12**(8), 3228 (2020)
8. Armstrong, M., Taylor, S.: *Armstrong's Handbook of Human Resource Management Practice: A Guide to the Theory and Practice of People Management*. Kogan Page Publishers (2023)

9. Fawehinmi, O., Yusliza, M.Y., Mohamad, Z., et al.: Assessing the green behaviour of academics: the role of green human resource management and environmental knowledge. *Int. J. Manpow.* **41**(7), 879–900 (2020)
10. Yusoff, Y.M., Nejati, M., Kee, D.M.H., et al.: Linking green human resource management practices to environmental performance in hotel industry. *Glob. Bus. Rev.* **21**(3), 663–680 (2020)



Innovative Application of Virtual Reality Technology in Corporate Culture Experience

Jing Wang^(✉) and Ning Yang

Shandong Business Institute, Yantai 264670, Shandong, China
214868731@qq.com

Abstract. With the rapid development of information technology, virtual reality technology, with its immersion, interaction and imagination, provides a new platform for the display, training and promotion of corporate culture. Firstly, this paper summarizes the basic characteristics of virtual reality technology and its application potential in enterprise culture. Virtual reality technology has shown a wide range of potential applications and far-reaching impact in many fields. This study aims to further explore its application and impact in cultural experience. This paper focuses on the analysis of the specific application of virtual reality technology in corporate culture display, training and promotion, and compares the changes of various indicators before and after the application in the form of tables, including employee satisfaction, training efficiency, employee retention, recruitment efficiency and brand awareness. The results show that the application of virtual reality technology significantly improves these indicators and enhances the corporate culture experience of employees.

Keywords: Virtual Reality · VR · Enterprise Culture Experience · Cooperate Culture Experience

1 Introduction

Virtual reality technology provides users with immersive experience by simulating the real environment, which shows great potential and application value in many fields. In the field of corporate culture experience, the application of virtual reality technology is particularly noticeable, which can not only enhance employee's sense of identity with corporate culture, but also play an important role in recruitment, training, team building and so on. Corporate culture is the soul of an enterprise, which embodies the values, code of conduct and business philosophy of an enterprise, and has a far-reaching impact on the long-term development of an enterprise. Traditional ways of experiencing corporate culture, such as corporate history exhibitions, employee handbooks and training courses, often have some limitations, such as lack of intuitive information transmission, low participation and weak sense of experience. The application of virtual reality technology provides a new perspective and method for corporate culture experience. Virtual reality technology can provide an immersive experience environment, so that employees can feel the development process, core values and business philosophy of the

enterprise as if they were on the scene. Through virtual reality technology, employees can “walk into” the past of the enterprise, understand the history and cultural heritage of the enterprise, “participate” in the major events of the enterprise, feel the pulse of the development of the enterprise, “experience” the social responsibility and public welfare actions of the enterprise, and enhance the sense of social responsibility. This immersive experience is far more vivid and intuitive than traditional words and pictures, and can affect employee’s emotions and cognition more profoundly. Through the systematic research on the application of virtual reality technology in corporate culture experience, this paper aims to provide a new perspective for enterprises to help them make better use of virtual reality technology and enhance the dissemination effect and influence of corporate culture.

2 Literature Review

Asatiani et al. (2021) explores the construction of organizational culture continuity in highly virtual work environments, providing insights for understanding virtual organizational culture [1]. Krajčovič et al. (2021) studied the application of virtual reality educational games in lean management teaching, demonstrating its teaching potential [2]. Kumari et al. (2022) discussed the application prospect of virtual reality technology in the field of agriculture, which is of great significance to improve the quality of agriculture [3]. Trushkina et al. (2020) analyzed the digital transformation of organizational culture under the condition of information economy, which provided a new perspective for the study of organizational culture [4]. Dluhopolskyi et al. (2021) explored the economic potential of virtual reality in the digital society, highlighting its importance in the economic field [5]. Carter & Egliston (2023) discussed the risks of VR data, including learning analytics and algorithmic biases, and criticized data perfectionism [6]. Akdere et al. (2022) evaluated virtual reality simulation training and explored the frontier issues of people and technology [7]. Garbin Prani Praničević (2021) studied the application of augmented reality and virtual reality technology in cultural tourism, providing a new experience for cultural tourism [8]. Nisiotis et al. (2020) proposed a prototype that integrates virtual reality, robotics and social networks, aiming to create a new social ecosystem of cultural heritage [9]. Guo et al. (2020) systematically reviewed the application of virtual reality in the life cycle maintenance of industrial products, providing a reference for industrial maintenance [10]. To sum up, virtual reality technology has shown a wide range of potential applications and far-reaching impact in many fields, and this paper aims to further explore its application and impact in specific areas.

3 Virtual Reality Technology and Corporate Culture

3.1 Virtual Reality Technology

Virtual reality is that the feeling in the virtual environment is similar to the feeling in reality, even directly approaching the real feeling, and its verisimilitude and real-time interaction provide strong support for the system simulation technology. Virtual reality has the following three basic characteristics, namely immersion, interaction and

imagination. The virtual reality system is mainly composed of a software platform, a virtual three-dimensional model database, an input device and an output device, wherein the input signal of the system mainly comes from different posture information of a user, or the input signal is obtained by collecting other information of a human body through other special data software. Virtual software Unity3D software converts the input information into the characters and scene information in the virtual scene to update the scene in the user's vision, and transmits its important output information (such as collision and sound of collision detection) to the output device. When the user obtains the output information, he can analyze it and then make the next action. AR is the real world plus virtual information, virtual reality is a completely virtual world, and MR is the combination of the real world and the virtual world, and the two can influence each other. The relationship between the three is shown in Fig. 1.

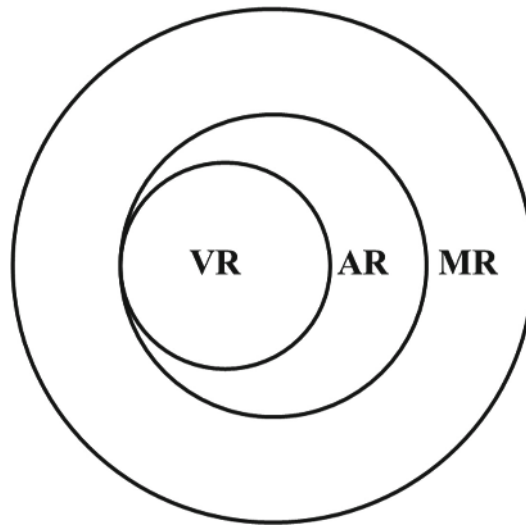


Fig. 1. Relationship among AR, MR and VR

Virtual reality industry is a diversified ecosystem, covering all aspects from core hardware to software systems, interactive technology, content services and head-mounted display devices. In this industry, large Internet enterprises, through investment or independent research and development, lock in potential enterprises and cutting-edge technologies, carry out industry layout, and use their own resources and capabilities to lead the development of the industry. Professional virtual reality enterprises, as the main force in the head display market, focus on the development of display technology and related software and hardware, and constantly introduce new products.

In 2021, due to the rapid introduction of new products and diversified marketing strategies by manufacturers, as well as the popularity of the concept of “meta-universe”, virtual reality devices were regarded as a passport to enter the meta-universe, and global shipments exceeded 10 million units for the first time. However, with the overall cooling of the technology industry, the leading companies in the virtual reality industry have not

announced positive market news in the past year. As shown in Fig. 2, 7.65 million virtual reality devices were shipped worldwide in 2023, with Meta accounting for more than 70% of the market share. In the post-epidemic era, people’s leisure and entertainment choices have become more diversified, and virtual reality devices whose performance is not yet fully mature are difficult to attract a large number of users. Although the launch of Vision Pro at the beginning of the year has played a certain role in promoting the XR market, manufacturers still need time to improve the hardware and content ecosystem. As a result, global VR device shipments are expected to grow slightly to more than 8.1 million units in 2024.

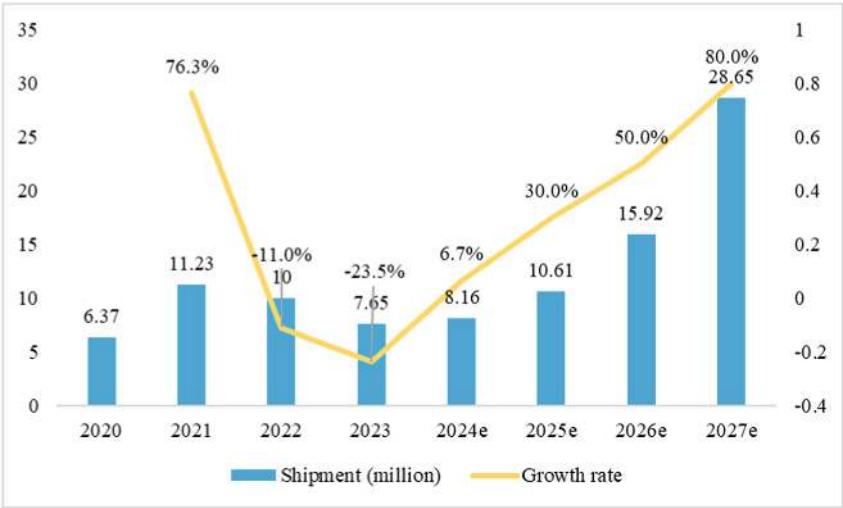


Fig. 2. Shipment volume and forecast of global VR terminal equipment

3.2 Corporate Culture

Corporate culture is essentially the combination of all kinds of tangible and intangible in the operation and development of the members of the enterprise. For the enterprise, it mainly involves the spirit of the enterprise, the enterprise system, the production environment and the development process. For all members of the enterprise, it mainly involves the common values, ethics and codes of conduct. Among them, the most critical is the values, the core content of corporate culture is actually the common decision of the values recognized by employees. Generally speaking, corporate culture includes its management philosophy, value, enterprise spirit, enterprise image, system and so on. Corporate material culture is a collection of visible and real products or services, external material characteristics and so on. The common behavior characteristics or management habits of enterprise members in production and life can be summarized as the behavior culture of enterprises. Enterprise system culture is the way of overall management of enterprises, the main purpose of which is to restrict and standardize the

behavior of employees. For the overall management of enterprises, the enterprise system is composed of decision-making system, organizational structure and so on. Corporate spiritual culture includes the vision, values and mission of an enterprise, which is a kind of group consciousness at the spiritual level different from material culture.

The guiding function of corporate culture enables employees to actively control their own behavior according to the strategic planning, rules and regulations of the enterprise. Good corporate culture can effectively improve the enthusiasm of enterprise members to work, so that they can maintain a positive and enterprising spirit. Many enterprises are accustomed to using material incentives for employees, such as salary, welfare and other incentives, but such incentives usually can only play a short-term role, and cannot achieve long-term effective incentives. In contrast, spiritual motivation can achieve a more effective role and play a longer role. Companies with good corporate culture can bring better working atmosphere for employees, so that employees can participate in work more happily. In addition, enterprises will pay more attention to the needs of employee's personal feelings and self-worth realization, which will effectively enhance the enthusiasm of employees for work.

4 Innovative Application of Virtual Reality Technology in Corporate Culture Experience

4.1 Display of Corporate Culture

Through virtual reality technology, employees and customers can experience the core values and cultural atmosphere of the enterprise, interact with the leaders of the enterprise in the virtual world, and understand the mission and vision of the enterprise. This immersive experience can not only help employees better integrate into the corporate culture, but also enhance their sense of belonging and identity, and enhance their work enthusiasm and creativity. Virtual reality technology can also help enterprises better display their corporate philosophy and cultural characteristics. Through virtual reality technology, enterprises can produce various virtual reality display contents, such as enterprise history review, enterprise mission statement, employee story sharing, etc., so that employees and customers can more intuitively understand the development process, core concepts and cultural connotations of enterprises. This kind of display can not only enhance the corporate image, but also enhance the emotional connection between enterprises and employees and customers, and promote the loyalty and recognition of employees and customers. Through virtual reality technology, enterprises can create a virtual cultural space, in which employees can conduct virtual training and interactive experience, and cultivate their sense of teamwork, innovation and leadership. At the same time, enterprises can also use virtual reality technology to record and inherit the essence of corporate culture, so that corporate culture can continue and develop, laying a solid foundation for the sustainable development of enterprises. Virtual reality technology can also present the core values and vision of corporate culture in a vivid and intuitive way. Through the construction of virtual scenes, it shows the practice and achievements of enterprises in social responsibility, innovative spirit, teamwork and other aspects, so that abstract ideas become concrete and sensible.

4.2 Training of Corporate Culture

Virtual reality technology can create highly realistic virtual scenes, so that employees feel as if they are in a real working environment. For example, by simulating the company's office, production line or customer service scenarios, employees can experience and practice daily work processes in a virtual environment, and better understand the embodiment of corporate culture in real work. In terms of training content delivery, virtual reality technology can present corporate culture in a more vivid and interesting way. It is no longer a boring text explanation, but through the interaction of virtual characters and the unfolding of story plots, employees can deeply feel the mission, vision and values of the enterprise. This kind of immersive experience can arouse the emotional resonance of employees and make them accept and identify with the corporate culture more actively. Virtual reality technology also supports multi-person online interaction and promotes collaboration and communication among employees. In the virtual training scenario, employees can complete tasks in groups and solve problems together, thus cultivating team spirit, which is in line with the concept of solidarity and cooperation emphasized in corporate culture. Virtual reality technology can provide personalized training content and feedback according to the different needs and learning progress of employees. It can monitor the performance of employees in the virtual environment in real time, provide targeted suggestions and improvement directions for employees, and improve the training effect. Corporate culture training often involves some complex or difficult to simulate scenarios. Virtual reality technology can easily realize the simulation of these scenes, such as emergency evacuation drills, customer service simulation and so on. Employees can learn how to deal with all kinds of complex situations and improve their ability to solve problems without actual risks. Virtual reality technology can provide personalized learning paths according to the learning progress and ability of employees. Employees can learn according to their own rhythm, and the system can adjust the training content and difficulty according to the performance of employees, so as to ensure that each employee can get the most suitable training experience.

4.3 Promotion of Corporate Culture

Corporate culture promotion is not only a one-way information transmission, but also a two-way interactive exchange. Virtual reality technology allows users to interact with corporate culture elements in a virtual environment, such as participating in virtual product experience and dialogue with virtual corporate representatives. This kind of interaction can enhance the user's recognition and loyalty to the brand. In the process of recruitment, enterprises can use virtual reality technology to show potential candidates the working environment and cultural atmosphere of enterprises, and attract their interest and willingness to join. This novel recruitment method can improve the attractiveness of enterprises and help enterprises gain an advantage in the talent competition. Enterprises can use virtual reality technology to promote the market, such as through virtual reality experience to show the functions and advantages of products, or through virtual activities to enhance interaction with consumers. This innovative means of promotion can enhance the attractiveness and influence of market activities. For multinational enterprises, virtual

reality technology can help employees with different cultural backgrounds better understand and integrate into corporate culture. Through virtual cross-cultural experience, employees can enhance respect and understanding of different cultures, and promote the diversity and inclusiveness of corporate culture. Although virtual reality technology shows great potential in the promotion of corporate culture, it also faces challenges such as technology cost, content development and user acceptance. Enterprises need to consider these factors comprehensively and formulate reasonable promotion strategies to ensure that virtual reality technology can effectively serve the promotion of corporate culture.

5 Application Effect of Virtual Reality Technology in Corporate Culture Experience

As shown in Table 1, virtual reality technology can significantly improve employees' perception and satisfaction of corporate culture by providing immersive experience. Compared with traditional training methods, virtual reality technology can more vividly convey corporate values and mission, thereby enhancing the sense of belonging and identity of employees. The use of virtual reality for training can simulate a variety of work scenarios, so that employees can practice in a safe environment, which not only improves the safety of training, but also improves the efficiency and effectiveness of training. Enhancing the experience of corporate culture through virtual reality technology can enhance employee loyalty and retention. In-depth understanding and recognition of corporate culture will help to reduce employee turnover. The application of virtual reality technology in the recruitment process can show potential candidates the working environment and culture of enterprises, attract more attention of outstanding talents, and improve the efficiency and quality of recruitment. The promotion of corporate culture through virtual reality technology can enhance the public's awareness of corporate brand and enhance brand image.

Table 1. The effect of virtual reality technology application in corporate culture experience

Indicators	Before VR application	After VR application	Rate of change
Employee satisfaction	7.5	8.8	17.33%
Training efficiency	4.5	7.2	60.00%
Employee retention	65%	75%	14.07%
Recruitment efficiency	5.2	8.1	55.77%
Brand recognition	6.2	7.9	27.42%

As can be seen from the table, employee satisfaction has increased from 7.5 to 8.8, with a change rate of 17.33%. This shows that the application of virtual reality technology has significantly improved employees' perception and satisfaction of corporate culture. Immersive experiences may give employees a deeper understanding of the values and mission of the enterprise, thus enhancing their sense of belonging and identity.

The improvement of training efficiency is particularly significant, from 4.5 to 7.2, with a change rate of 60.00%. This may be attributed to the highly simulated environment of virtual reality technology, which enables employees to practice in a safe and controlled environment, thus acquiring the necessary skills and knowledge more quickly. The improvement of employee retention rate shows that the corporate culture experience enhanced by virtual reality technology can help to promote employee loyalty. Employees' in-depth understanding and recognition of corporate culture will help to reduce the turnover rate of employees, which is essential for the long-term development of enterprises. The increase in recruitment efficiency (from 5.2 to 8.1) shows the great potential of virtual reality technology in attracting potential talents. By displaying the working environment and culture of enterprises through virtual reality, enterprises can attract excellent talents more effectively and improve the quality of recruitment. The improvement of brand recognition (from 6.2 to 7.9) reflects the effect of virtual reality technology in the promotion of corporate culture. This innovative way of promotion not only improves the public's awareness of the corporate brand, but also enhances the brand image.

On the whole, the application of virtual reality technology in corporate culture experience has brought comprehensive and positive effects. Especially, the improvement of training efficiency and recruitment efficiency shows that virtual reality technology can effectively optimize the human resource management process of enterprises. At the same time, the promotion of employee satisfaction and brand awareness also shows the potential of virtual reality technology in enhancing the soft power of corporate culture.

6 Conclusions

The application of virtual reality technology in corporate culture experience not only improves employee satisfaction and retention rate, but also significantly improves training efficiency and recruitment efficiency, while enhancing corporate brand awareness. These achievements fully prove the great potential and value of virtual reality technology in the construction of enterprise culture. However, we must also recognize that in the process of promoting virtual reality technology, enterprises need to face challenges such as technology cost, content development and user acceptance. Therefore, enterprises should consider these factors comprehensively and formulate reasonable strategies to ensure the effective application of virtual reality technology. In addition, enterprises should also pay attention to the latest development of virtual reality technology in order to adopt new technologies in time and maintain competitive advantage.

Acknowledgements. Construction Plan for Corporate Culture Core Concept of Yantai Zhuoming Property Management Co., Ltd. (Horizontal Project of Shandong Business Vocational College, No. 2024HXZX013).

References

1. Asatiani, A., Hämäläinen, J., Penttinen, E., et al.: Constructing continuity across the organisational culture boundary in a highly virtual work environment. *Inf. Syst. J.* **31**(1), 62–93 (2021)

2. Krajčovič, M., Gabajová, G., Furmannová, B., et al.: A case study of educational games in virtual reality as a teaching method of lean management. *Electronics* **10**(7), 838 (2021)
3. Kumari, S., Raghuram, P., Venkatesh, V.G., et al.: Future perspectives on progressive farming with adoption of virtual reality technology for sustainable quality in agriculture. *TQM J.* **34**(2), 250–279 (2022)
4. Trushkina, N., Abazov, R., Rynkevych, N., et al.: Digital transformation of organizational culture under conditions of the information economy. *Virtual Econ.* **3**(1), 7–38 (2020)
5. Dluhopolskyi, O., Simakhova, A., Zatonatska, T., et al.: Potential of virtual reality in the current digital society: economic perspectives. In: 2021 11th International Conference on Advanced Computer Information Technologies (ACIT). IEEE, pp. 360–363 (2021)
6. Carter, M., Egliston, B.: What are the risks of virtual reality data? Learning analytics, algorithmic bias and a fantasy of perfect data. *New Media Soc.* **25**(3), 485–504 (2023)
7. Akdere, M., Jiang, Y., Lobo, F.D.: Evaluation and assessment of virtual reality-based simulated training: exploring the human–technology Frontier. *Eur. J. Train. Dev.* **46**(5/6), 434–449 (2022)
8. Garbin, P.D.: Augmented reality and virtual reality-based technology in cultural tourism[J]. *ENTRENOVA-ENTERprise REsearch InNOVation* **7**(1), 307–314 (2021)
9. Nisiotis, L., Alboul, L., Beer, M.: A prototype that fuses virtual reality, robots, and social networks to create a new cyber–physical–social eco-society system for cultural heritage. *Sustainability* **12**(2), 645 (2020)
10. Guo, Z., Zhou, D., Zhou, Q., et al.: Applications of virtual reality in maintenance during the industrial product lifecycle: a systematic review. *J. Manuf. Syst.* **56**, 525–538 (2020)



Analysis on the Influencing Factors of Dual-Employee Family Work Engagement Based on Resource Conservation Theory: A Computer Aided Study

Xiaorui Shi^{1,2} and Yiwen Chen^{1,2(✉)}

¹ CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China
chenyw@psych.ac.cn

² Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049,
China

Abstract. The purpose of this study is to explore the influencing factors of dual-employee family work engagement. On the research method, the theoretical model is constructed based on the resource conservation theory, and the computer-aided research method is adopted, combined with questionnaire survey and structural equation model analysis for empirical research. The results show that work resources (including work autonomy, promotion opportunities and social support) have a significant positive impact on the work engagement of dual-earner families. At the same time, family resources (such as family support, family atmosphere and parent-child relationship) also indirectly support job engagement. Work-family conflict plays an intermediary role between work resources and work engagement, indicating that the shortage of work resources may aggravate work-family conflict and then reduce work engagement. The study also verified the existence of resource spiral effect, that is, there is a significant mutual promotion relationship between work input and the growth of family resources. These research results provide a new perspective for understanding the complex mechanism of dual-employee family work engagement.

Keywords: Dual-employee Family · Work Input · Resource Conservation Theory · Work-family Conflict · Computer Aided

1 Introduction

In today's society, dual-employee family has become a common family model. With the promotion of women's professional status and the increase of family economic pressure, more and more families choose both husband and wife to devote themselves to the workplace and share the family's economic responsibility [1]. However, this family model has also brought new challenges and problems, especially the impact on work engagement. As the embodiment of employees' recognition, enthusiasm and hard work, job engagement is not only related to personal career development, but also has a far-reaching impact on the overall performance and competitiveness of the organization

[2]. The balance between work and family faced by dual-income families is particularly prominent [3]. Because both husband and wife need to devote time and energy to their work, this often leads to conflicts and pressures between their family roles and work roles [4]. This kind of conflict and pressure may further affect their work engagement, and even have a negative impact on their career and family life. Therefore, studying the influencing factors of dual-employee families' job involvement will not only help reveal the special challenges faced by this group, but also provide theoretical basis for formulating effective intervention measures.

The resource conservation theory provides a theoretical framework to understand the influencing factors of dual-employee family work engagement. This theory emphasizes the individual's protection motivation in the face of resource loss and the value-added motivation in resource acquisition [5]. In a dual-employee family, work resources and family resources are two important forms of resources. Work resources include job promotion, work autonomy and social support, while family resources involve family support, family atmosphere and parent-child relationship [6]. According to the theory of resource conservation, when dual-income families face resource loss, they will take a series of protective actions to prevent further losses; And when they get resources, they will seek more value-added opportunities [7].

Although the theory of resource conservation has achieved rich research results in explaining employees' job involvement, its application in the specific group of dual-employee families is still relatively limited [8]. In addition, most of the existing research on dual-income family work engagement adopts traditional questionnaire survey and interview methods, which have certain limitations in data collection, management and analysis [9]. Computer-aided research can improve the efficiency and accuracy of data collection, and can also analyze a large number of data in depth through advanced statistical software [10]. Based on the theory of resource conservation, this study aims to explore the influencing factors of dual-employee family work engagement, and adopt computer-aided research methods to improve the efficiency and accuracy of the study. Specifically, the study will build a theoretical model including variables such as work resources, family resources and work-family conflict, collect the data of dual-employee families through questionnaires, and analyze the structural equation model with Amos software. Through this research design, it is expected to reveal the internal mechanism of dual-employee family work investment and provide theoretical basis and practical guidance for formulating effective intervention measures.

2 Theoretical Basis and Research Hypothesis

2.1 Theoretical Basis: Resource Conservation Theory

As the theoretical basis of this study, resource conservation theory provides a powerful framework for understanding the influencing factors of dual-employee family work engagement. Its core view is that individuals are always actively trying to maintain, protect and build their own limited resources [11]. These resources can be material, psychological, social or organizational, which are very important for the survival, development and happiness of individuals [12]. The theory of resource conservation includes two basic principles: the principle of resource loss and the principle of resource investment.

① Principle of resource loss: This principle points out that when individuals are faced with resource loss, they will have strong protection motivation to avoid further losses [13]. The loss of resources is not limited to material reduction, but also includes the consumption of intangible resources such as psychological energy and social support [14]. In a dual-employee family, the reduction of work resources or the loss of family resources may lead to the individual's resource protection behavior, which in turn affects his work input.

② Principle of resource investment: Corresponding to the principle of resource loss, the principle of resource investment emphasizes that individuals will take positive investment behavior in order to obtain and maintain resources [15]. This kind of investment behavior can be direct (such as improving vocational skills through learning) or indirect (such as gaining social support through establishing good interpersonal relationships). In a dual-employee family, individuals may seek more work resources (such as promotion opportunities and salary increase) by increasing work input, or enhance the supporting role of family resources by improving family relations.

The theory of resource conservation also emphasizes the spiral effect of resources, that is, the acquisition of resources will promote more resource acquisition, while the loss of resources may lead to more resource loss. This effect is particularly significant in dual-income families, because the resources in the two fields of work and family are interrelated, and the changes in resources in one field often spread to the other.

2.2 Research Hypothesis

Based on the theory of resource conservation, combined with the particularity of dual-income families and the influencing factors of work engagement, this study puts forward the following assumptions:

① The positive influence of work resources on work engagement.

Hypothesis 1: The working resources (such as work autonomy, promotion opportunities, social support, etc.) of dual-employee families have a positive impact on their work engagement. The richness of work resources can meet the individual's demand for work and improve their job satisfaction and engagement.

② The supporting role of family resources in job engagement.

Hypothesis 2: Family resources (such as family support, family atmosphere, parent-child relationship, etc.) of dual-employee families indirectly support their work engagement. Sufficient family resources can relieve the pressure brought by work, improve the individual's psychological well-being, and thus enhance their work engagement.

③ Mediating effect of work-family conflict.

Hypothesis 3: Work-family conflict plays an intermediary role between work resources and work engagement in dual-employee families. The shortage of work resources may aggravate the work-family conflict, and then affect the individual's work engagement. On the contrary, rich family resources may help to alleviate work-family conflicts and promote work engagement.

④ Verification of resource spiral effect.

Hypothesis 4: There is a spiral effect between job involvement and the growth of family resources in dual-employee families. That is to say, the increase of work input will promote the growth of family resources (such as improving family atmosphere and

strengthening parent-child relationship), and the growth of family resources will further promote the increase of work input.

In order to further verify these hypotheses, this study will construct a structural equation model (SEM) including variables such as work resources, family resources, work-family conflict and work engagement. The data of dual-income families were collected through questionnaire survey, and the model fitting and path analysis were carried out by using Amos software.

3 Research Method

3.1 Research Objects and Samples

The research object of this study is the dual-employee family, that is, the family model in which both husband and wife work in the workplace. In order to ensure the representativeness and effectiveness of the study, we selected dual-income families from different industries, different positions and different working years as research samples by random sampling. The selection of samples should fully consider multiple factors such as region, age, gender and educational background, so as to ensure the universality and popularization of the research results.

3.2 Data Collection

In the data sorting stage, the questionnaire data are coded and entered to ensure the accuracy and consistency of the data. In the stage of data cleaning, abnormal values and missing values are handled to ensure the integrity and reliability of data. In the data analysis stage, SEM is used for in-depth statistical analysis. Through SEM analysis, we can verify the applicability of the resource conservation theory in the work engagement of dual-employee families, and discuss the specific influence paths and effects of work resources, family resources and work-family conflict on work engagement.

Data acquisition technology: Computer-aided online questionnaire system and professional online survey platform are used for data acquisition. The system has the functions of automatic questionnaire distribution, data collection and preliminary sorting, and can monitor the progress of questionnaire collection in real time, and preliminarily clean and screen the data to eliminate invalid questionnaires.

Data analysis tool: structural equation model (SEM) analysis was carried out by Amos software. Amos software can deal with complex causality models, estimate model parameters by means of maximum likelihood estimation, and provide model fitting indicators to evaluate the goodness of fit of the model. In addition, Amos also supports multiple groups of comparative analysis, which can be used to explore the differences between different groups (such as different industries and different sexes).

Data preprocessing technology: in the data preprocessing stage, the missing data is processed by missing value interpolation technology to reduce the influence of data loss on the analysis results. At the same time, outlier detection algorithm is used to identify and process outliers to ensure the reliability of data.

Model verification technology: cross-validation method is used to verify the structural equation model to evaluate its stability and generalization ability. In addition, the

confidence interval of model parameters is estimated by Bootstrap method, which further verifies the reliability of model results.

3.3 Data Management and Analysis

In the data sorting stage, the questionnaire data are coded and entered to ensure the accuracy and consistency of the data. In the data cleaning stage, abnormal values and missing values are processed to ensure the integrity and reliability of data. In the data analysis stage, structural equation model (SEM) is used for in-depth statistical analysis. Through SEM analysis, we can verify the applicability of the resource conservation theory in the work engagement of dual-employee families, and discuss the specific influence paths and effects of work resources, family resources and work-family conflict on work engagement.

3.4 Study Quality Control

In order to ensure the scientificity and accuracy of the research, a series of measures are taken to control the research quality. First of all, in the questionnaire design stage, pre-test and revision are carried out to ensure the accuracy and pertinence of the questionnaire questions. Secondly, in the stage of data collection, strict screening and auditing are carried out on the families with dual employees who participate in the survey to ensure the representativeness and validity of the samples. Finally, in the data analysis stage, various statistical methods are used to verify the stability and reliability of the research results.

4 Research Results

4.1 Influence of Work Resources on Work Engagement

Table 1 shows the analysis results of the influence of job resources on job engagement. As can be seen from the table, work resources such as work autonomy, promotion opportunities and social support have a significant positive impact on job engagement. Specifically, for each additional unit of work autonomy, the work input will increase by 0.32 units; For every promotion opportunity, the job input will increase by 0.27 units; For every additional unit of social support, the work input will increase by 0.41 unit. These data show that rich work resources can effectively improve the work input of dual-income families.

4.2 Supporting Role of Family Resources in Job Engagement

Table 2 shows the analysis results of the supporting role of family resources in job engagement. Family resources, such as family support, family atmosphere and parent-child relationship, have indirect support for job engagement. For each additional unit of family support, the work input is indirectly increased by 0.18 units by alleviating the work-family conflict; Every time the family atmosphere is improved by one level, the

Table 1. Influence of work resources on work engagement

Work Resources	Regression Coefficient	Standard Error	t-value	p-value
Job Autonomy	0.32	0.05	6.40	< 0.001
Promotion	0.27	0.04	6.75	< 0.001
Social Support	0.41	0.06	6.83	< 0.001

work input is indirectly increased by 0.15 units; For each unit of parent-child relationship, 0.22 units will be invested in indirect promotion. These data show that sufficient family resources can alleviate the pressure brought by work, and then enhance the work input of dual-earner families.

Table 2. Supporting role of family resources in job engagement

Family Resources	Indirect Effect	Standard Error	z-value	p-value
Family Support	0.18	0.03	6.00	< 0.001
Family Atmosphere	0.15	0.04	3.75	< 0.001
Parent-Child Relationship	0.22	0.05	4.40	< 0.001

4.3 Mediating Effect of Work-Family Conflict

Table 3 shows the analysis results of the mediating effect of work-family conflict between work resources and work engagement. Work-family conflict plays a partial mediating role between work resources such as work autonomy, promotion opportunities and social support and work engagement. The direct influence of work autonomy on work input is 0.20 units, and the indirect influence through work-family conflict is 0.12 units; The direct impact of promotion opportunities on job engagement is 0.18 units, and the indirect impact through work-family conflict is 0.09 units; The direct impact of social support on job engagement is 0.30 units, and the indirect impact through work-family conflict is 0.11 units. These data show that the shortage of work resources may aggravate the work-family conflict, and then affect the work input of dual-earner families.

Table 3. Mediation effect of work-family conflict

Work Resources	Direct Effect	Indirect Effect	Total Effect
Job Autonomy	0.20	0.12	0.32
Promotion	0.18	0.09	0.27
Social Support	0.30	0.11	0.41

4.4 Verification of Resource Spiral Effect

Table 4 shows the verification and analysis results of resource spiral effect. There is a significant spiral effect between job involvement and the growth of family resources in dual-employee families. For each additional unit of work input, family support will increase by 0.25 units, family atmosphere will improve by 0.18 units, and parent-child relationship will be enhanced by 0.30 units. On the contrary, the growth of family resources such as family support, family atmosphere and parent-child relationship will further promote the increase of work input. These data show that in dual-income families, work input and the growth of family resources promote each other and form a virtuous circle.

Table 4. Verification of resource spiral effect

Variables	Regression Coefficient	Standard Error	t-value	p-value
Work Engagement → Family Support	0.25	0.04	6.25	< 0.001
Work Engagement → Family Atmosphere	0.18	0.03	6.00	< 0.001
Work Engagement → Parent-Child Relationship	0.30	0.05	6.00	< 0.001

The results show that work resources, family resources, work-family conflict and resource spiral effect all have a significant impact on the work engagement of dual-earner families.

5 Discussion

5.1 The Importance of Work Resources for Work Engagement in Dual Income Families

The results show that work resources such as work autonomy, promotion opportunities and social support have a significant positive impact on the work engagement of dual-earner families. This discovery is consistent with the core view of resource conservation theory, that is, individuals tend to acquire and preserve resources to meet the challenges in work and life. In a dual-employee family, both husband and wife need to strike a balance between work and family, and rich work resources can provide them with more autonomy and control, and reduce work pressure.

Work autonomy enables dual employees to arrange working hours and tasks more flexibly, thus better coping with the conflict between family and work. Promotion opportunities provide the possibility of career development for dual employees, and enhance their work motivation and sense of belonging. Social support helps dual employees to establish positive interpersonal relationships in the workplace, get emotional satisfaction and support, and then enhance their work engagement.

Work resources can be regarded as an “input signal”, while work input is an “output response”. According to the system theory, a good input signal can optimize the output performance of the system. Work autonomy enables dual employees to arrange working hours and tasks more flexibly, thus better coping with the conflict between family and work. Promotion opportunities provide the possibility of career development for dual employees and enhance their work motivation and sense of belonging. This is similar to introducing incentive mechanism into the career development system to improve the efficiency of the system. Social support helps dual employees to establish positive interpersonal relationships in the workplace, get emotional satisfaction and support, and thus enhance their work engagement.

5.2 The Indirect Supporting Function of Family Resources to Work Engagement

The study also found that family resources such as family support, family atmosphere and parent-child relationship have indirect support for the work engagement of dual-earner families. This discovery emphasizes the important role of family as an important part of individual resources in work engagement. Family resources not only directly affect the individual’s mental health and quality of life, but also indirectly enhance work engagement by alleviating the work-family conflict.

In a dual-employee family, family support includes mutual understanding, support and cooperation between husband and wife, and family members’ recognition and respect for their work. This kind of support helps to reduce the work pressure of dual employees and enhance their job satisfaction and engagement. Family atmosphere reflects the emotional connection and interaction between family members. Parent-child relationship is an important part of family resources, and good parent-child relationship helps dual employees to better balance their work and family roles.

From the perspective of systems engineering, family resources can be regarded as an “auxiliary system”. It improves the individual’s performance in the work system by optimizing his psychological state. Family support includes mutual understanding, support and cooperation between husband and wife, and family members’ recognition and respect for their work. This kind of support helps to reduce the work pressure of dual employees and enhance job satisfaction and engagement, which is similar to the introduction of collaborative mechanism in multi-system interaction to improve the overall efficiency. Family atmosphere reflects the emotional connection and interaction between family members. A good family atmosphere can provide emotional comfort and motivation for dual employees, so that they can maintain a positive attitude in the face of work pressure. Parent-child relationship is an important part of family resources, and good parent-child relationship helps dual employees to better balance their work and family roles.

5.3 Mediating Effect of Work-Family Conflict

This study also reveals the mediating effect of work-family conflict between work resources and work engagement. This discovery shows that the shortage of work resources may aggravate the work-family conflict, and then affect the work engagement of dual-income families. This result further verifies the applicability of the resource

conservation theory and emphasizes the importance of striking a balance between work and family.

In a dual-earner family, work and family are two interrelated fields. The shortage of working resources may lead to more challenges and pressures faced by dual employees at work, which will further affect their role in the family. This kind of work-family conflict will have a negative impact on the mental health and quality of life of dual-employees, reducing their work input and career development potential.

Work-family conflict can be regarded as a “interference signal”, which will weaken the positive impact of work resources on work engagement. In a dual-earner family, work and family are two interrelated fields. The shortage of working resources may lead to more challenges faced by dual employees in their work. This kind of work-family conflict will have a negative impact on the mental health and quality of life of dual-employees, reducing their work input and career development potential. Therefore, reducing work-family conflict is the key link to improve the work engagement of dual employees.

5.4 Enlightenment of Spiral Effect of Resources

Finally, this study verifies the existence of resource spiral effect in dual-employee families. The results show that there is a significant spiral effect between job involvement and the growth of family resources, that is, the increase of job involvement will promote the growth of family resources, and the growth of family resources will further promote the increase of job involvement. This discovery emphasizes the mutual promotion between job involvement and family resources, and provides a theoretical basis for formulating effective intervention measures.

In a dual-employee family, work input and family resources are interdependent and mutually reinforcing. Dual-income workers need to pay attention to the improvement of work input to promote the growth of family resources; At the same time, we also need to pay attention to the investment and preservation of family resources to support the continuous increase of work investment. The existence of this spiral effect of resources provides a virtuous circle development model for dual-employee families, which helps them to achieve a better balance between work and family.

The spiral effect of resources can be regarded as a “positive feedback loop”. In a dual-employee family, job involvement and family resources are interdependent and mutually reinforcing. Dual-income workers need to pay attention to the promotion of work input to promote the growth of family resources, and also pay attention to the investment and preservation of family resources to support the continuous increase of work input. This spiral effect of resources provides a virtuous circle development model for dual-career families, which helps them achieve a better balance between work and family.

References

1. Jie, H, Dani, C.M., et al.: The influence of parents' work engagement of dual-employee families on junior high school students' academic engagement. *Psychol. Sci.* **42**(6), 1368–1374 (2019)

2. Xiaoquan, W., Chunfu, S., Chaoying, Y.: Commuting time model of dual-employee families considering multi-scale built-up environmental impact. *J. Transp. Eng. Inform.* **21**(3), 121–129 (2023)
3. Mei, F.: Study on the parent-child relationship of high school students with dual-income families in the new era and reasonable intervention plan. *College Entrance Examination* **7**, 115 (2020)
4. Cruz, J.P., Alquwez, N., Ejercito, B.: Work engagement of nurses and the influence of spiritual climate of hospitals: a cross-sectional study. *J. Nursing Manage.* **30**(1), 279–287 (2022)
5. Mubarak, N., Khan, J., Yasmin, R., et al.: The impact of a proactive personality on innovative work behavior: the role of work engagement and transformational leadership. *Leadersh. Org. Dev. J.* **42**(7), 989–1003 (2021)
6. Hao, H., Tian, Q., Jianing, M., et al.: The influence of organizational justice on female nurses' job involvement: the mediating role of psychological capital. *China Health Stat.* **36**(02), 213–215 (2019)
7. Han, W.: The dilemma and countermeasures of paternal participation in dual-employee families from the perspective of social support. *Love, Marriage Family* **32**, 0161–0162 (2022)
8. Yi, D., Lifan, J., Kunaviktikul, W.: Analysis of influencing factors of nurses' job engagement based on JD-R model and countermeasures. *Chinese J. Mod. Nurs.* **9**, 1107–1111 (2018)
9. Ling, Y., Wenli, W.: Analysis of the influencing factors of rural kindergarten teachers' job involvement. *Changjiang Seri.* **30**, 119–121 (2022)
10. Yu, J., He, R., Wang, H., et al.: Analysis of the influencing factors of rural physical education teachers' job involvement and research on the promotion strategies. *J. Lvliang Inst. Educ.* **038**(002), 42–47, 61 (2021)
11. Sujuan, Z., Jiange, Z., Wenjing, F., et al.: Path analysis of the status quo of emergency nurses' job involvement and its influencing factors. *Gen. Nurs.* **21**(18), 2567–2570 (2023)
12. Lijun, X., Xiaoxia, W., Liu, Y.: Factors affecting female employees' job involvement and solutions. *Talent Outlook* **4**, 34–35 (2019)
13. Bi, X., Ma, M., Jin, J.: Work engagement of young employees in state-owned enterprises and its influencing factors. *Youth Stud.* **2021**(6), 48–58+92
14. Xingyong, T., Yaya, P., Yulu, L., et al.: Research on the influencing factors of employee creativity based on innovative process input. *Logistics Manage.* **027**(012), 172–180 (2018)
15. Jialing, L.: Research on the influencing factors of hotel employees' job engagement in china context. *Manage. Technol. Small Medium-sized Enterprises* **02**, 17–18 (2018)

Author Index

B

Bai, Yuehua 33

C

Chen, Guangren 44
Chen, Jingao 23
Chen, Jingjing 318
Chen, Ke 277
Chen, Kenjie 397
Chen, Meishan 103
Chen, Nan 207
Chen, Qingle 91
Chen, Rongwen 363
Chen, Yiwen 434
Cheng, Zhuopei 133

D

Deng, Jing 3
Du, Meiling 307

F

Fan, Guiling 232
Fan, Xiaodi 343

G

Gao, Li 80
Guo, Ge 343
Guo, Yu 23

H

Han, Zhaoyun 13
Heng, Wang 256
Hou, Weiling 415
Hu, Xiaoxu 23
Hu, Zhenyan 331
Huang, Yufei 103
Huang, Zhijie 103

J

Jiang, Chuyuan 397
Jiang, Min 57

K

Kong, Fanling 244, 406

L

Li, Chunjie 217
Liu, Weiyi 123
Liui, Lushi 13
Luo, Ruodan 277

M

Ma, Jiaqi 353
Ma, Yunjun 23
Meng, Shuyan 406

Q

Qin, Wenle 207
Qu, Zhifeng 181

R

Ruoyu, Xu 256

S

Shi, Xiaorui 434
Song, Meichao 80
Song, Yuanpeng 156
Sun, Nannan 114

T

Tang, Ruiqin 385
Tian, Guixian 318
Tian, Juanjuan 385

W

Wang, Chao 277
Wang, Huijun 143
Wang, Jing 425
Wang, Shuqing 265
Wang, Zhuqi 133
Wen, Jinfeng 217
Wen, Zhengxuan 193
Wu, Yi Ling 207
Wu, Yiling 13

X

Xie, Lili 415

Y

Yang, Diansheng 156
Yang, Haoyun 103
Yang, Lihua 298

Yang, Ning 425
Yang, Yu 33
Yanni, Xu 256
Ye, Rongrong 171

Z

Zhang, Dayong 363
Zhang, Peng 23
Zhang, Qinghe 133
Zhang, Qiwen 244
Zhang, Shaohui 397
Zhang, Zhihua 353
Zhang, Ziyu 288
Zhao, Ming 307
Zheng, Guogang 67
Zhou, Li 374
Zhou, Qiongjie 331