Smart Innovation, Systems and Technologies 411

Ireneusz Czarnowski
Robert J. Howlett
Lakhmi C. Jain *Editors*

Intelligent Decision Technologies

Proceedings of the 16th KES-IDT 2024 Conference

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 411

Series Editors

Robert J. Howlett, KES International, Shoreham-by-Sea, UK

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

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

Ireneusz Czarnowski · Robert J. Howlett ·
Lakhmi C. Jain
Editors

Intelligent Decision Technologies

Proceedings of the 16th KES-IDT 2024 Conference

Springer

Editors
Ireneusz Czarnowski
Gdynia Maritime University
Gdynia, Poland

Robert J. Howlett
KES International Research
Shoreham-by-Sea, UK

Lakhmi C. Jain
University of Piraeus
Athens, Greece

KES International
York, UK

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-981-97-7418-0 ISBN 978-981-97-7419-7 (eBook)
https://doi.org/10.1007/978-981-97-7419-7

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 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 Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Preface

This volume contains the proceedings of the 16th International KES Conference on Intelligent Decision Technologies (KES-IDT 2024). The conference was held in Santa Cruz, Madeira, Portugal, from 19 to 21 June 2024.

The KES-IDT is an international annual conference organized by KES International. The KES Conference on Intelligent Decision Technologies belongs to a sub-series of the KES Conference series.

The KES-IDT provides space for the presentation and discussion of new research results under the common title *Intelligent Decision Technologies*. The conference has an interdisciplinary character, giving opportunities to researchers from different scientific areas and different application areas to show how intelligent methods and tools can support the decision-making processes.

This year the submitted papers have been allocated to the main track and six special sessions. Each submitted paper has been reviewed by 2–3 members of the International Program Committee and International Reviewer Board. Thirty-nine papers were accepted for presentation during the conference and inclusion in the KES-IDT 2024 proceedings.

We are very satisfied with the quality of the papers and would like to thank the authors for choosing KES-IDT as the forum for the presentation of their work. We also would like to thank for interesting and valuable keynote speech during the conference.

We also gratefully acknowledge the hard work of the KES-IDT international program committee members and the additional reviewers for taking the time to review the submitted papers and selecting the best among them for presentation at the conference and inclusion in the proceedings.

We hope that KES-IDT 2024 significantly contributes to the fulfilment of academic excellence and leads to even greater successes of KES-IDT events in the future.

Gdynia, Poland — Ireneusz Czarnowski
Shoreham-by-Sea, UK — Robert J. Howlett
Athens, Greece — Lakhmi C. Jain
June 2024

KES-IDT 2024 Conference Organization

Honorary Chairs

Lakhmi C. Jain, KES International, UK & Liverpool Hope University, UK
Gloria Wren-Phillips, Loyola University, USA

General Chair

Ireneusz Czarnowski, Gdynia Maritime University, Poland

Executive Chair

Robert J. Howlett, KES International, UK & 'Aurel Vlaicu' University of Arad, Romania

Program Chairs

Jose L. Salmeron, University Pablo de Olavide, Seville, Spain
Antonio J. Tallón-Ballesteros, University of Huelva, Spain

Publicity Chairs

Izabela Wierzbowska, Gdynia Maritime University, Poland
Alfonso Mateos Caballero, Universidad Politécnica de Madrid, Spain

Keynote Speeches

Iván Macía, Department of Digital Health and Biomedical Technologies, Vicomtech, Spain
Hiroshi Takahashi, Graduate School of Business Administration, Keio University, Japan
Roman Sperka, Silesian University, School of Business Administration in Karvina, Czech Republic
Ivan Luković, University of Belgrade, Serbia

Special Sessions

Responsible and Trustworthy Artificial Intelligence
Alessia Amelio, Telematic University "Leonardo da Vinci, Italy
Daniela Cardone, Telematic University "Leonardo da Vinci, Italy
Arcangelo Merla, Telematic University "Leonardo da Vinci, Italy
Giampiero Di Plinio, Telematic University "Leonardo da Vinci, Italy
Andrea Monti, Telematic University "Leonardo da Vinci, Italy
Franco Sivilli, Telematic University "Leonardo da Vinci, Italy

Reasoning-based Intelligent Applied Systems
Jair Minoro Abe, Paulista University, Brazil

Statistical Analysis and Model Selection for Complex Structed Data
Rei Monden, Hiroshima University, Japan
Mariko Yamamura, Radiation Effects Research Foundation, Japan

Decision Making Theory for Economics
Takao Ohya, Kokushikan University, Japan
Takafumi Mizuno, Meijo University, Japan
Shunei Norikumo, Doshisha University, Japan

Large-Scale Systems for Intelligent Decision Making and Knowledge Engineering
Sergey V. Zykov, MEPhI National Nuclear Research University, Russia
Hadi M. Saleh, HSE University, Russia

Artificial Intelligence, Machine Learning and Deep Learning in Intelligent Decision Technologies
Mihaela Luca, Romanian Academy, Iaşi Branch, Romania
Anca Ignat, University "Alexandru Ioan Cuza" of Iaşi, Romania

International Program Committee and Reviewers

Ahmad Taher Azar
Valentina Emilia Balas
Dariusz Barbucha
Monica Bianchini
Adriana Burlea-Schiopoiu
Gloria Cerasela Crisan
Amitava Chatterjee
Matteo Cristani
Ireneusz Czarnowski
Vladimir Dimitrieski
Dinu Dragan
Agnieszka Duraj
Dawn E. Holmes
Margarita Favorskaya
Mauro Gaggero
Christos Grecos
Katsuhiro Honda
Tzung-Pei Hong
Dragan Ivanović
Yuji Iwahori
Dragan Jevtic
Nikos Karacapilidis
Pawel Kasprowski
Aleksandar Kovačević
Marek Kretowski
Vladimir Kurbalija
Pei-Chun Lin
Michele Mastroianni
Alfonso Mateos Caballero
Lyudmila Mila Mihaylova
Jim Prentzas
Marcos Quiles
Ana Respício
Gerasimos Rigatos
Hadi Saleh
Mika Sato-Ilic

Miloš Savić
Marek Sikora
Bharat Singh
Urszula Stańczyk
Jair Minoro Abe
Suneeta Mohanty
Mikhail Moshkov
Takao Ohya
Jeng-Shyang Pan
Camelia Pintea
Radu-Emil Precup
Catalin Stoean
Ruxandra Stoean
Piotr Szymczyk
Masakazu Takahashi
Shing Chiang Tan
Jan Treur
Eiji Uchino
Junzo Watada
Izabela Wierzbowska
Yoshiyuki Yabuuchi
Mariko Yamamura
Hiroyuki Yoshida
Gian Pierro Zarri
Beata Zielosko
Alfred Zimmermann
Sergey Zykov
Anca Ignat
Mihaela Luca
Takao Ohya
Hadi Saleh
Sergey Zykov
Ahmad Taher Azar
Jair Minoro Abe
Alessia Amelio
Daniela Cardone
Rei Monden
Mariko Yamamura
Elfazziki Abdelaziz
Milan Simic
Dariusz Barbucha
Ireneusz Czarnowski
Piotr Jędrzejowicz
Alexandru Tugui

Contents

About the Editors

Ireneusz Czarnowski is Professor at the Gdynia Maritime University. He holds B.Sc. and M.Sc. degrees in Electronics and Communication Systems from the same University. He gained the doctoral degree in the field of computer science in 2004 at Faculty of Computer Science and Management of Poznan University of Technology. In 2012, he earned a postdoctoral degree in the field of computer science in technical sciences at Wroclaw University of Science and Technology. His research interests include artificial intelligence, machine learning, evolutionary computations, multi-agent systems, data mining, and data science. He is Associate Editor of the Journal of Knowledge-Based and Intelligent Engineering Systems, published by the IOS Press, and Reviewer for several scientific journals.

Dr. Robert J. Howlett is Executive Chair of KES International, a non-profit organization that facilitates knowledge transfer and the dissemination of research results in areas including Intelligent Systems, Sustainability, and Knowledge Transfer. He is Visiting Professor at Bournemouth University in the UK. His technical expertise is in the use of intelligent systems to solve industrial problems. He has been successful in applying artificial intelligence, machine learning, and related technologies to sustainability and renewable energy systems; condition monitoring, diagnostic tools, and systems; and automotive electronics and engine management systems. His current research work is focused on the use of smart microgrids to achieve reduced energy costs and lower carbon emissions in areas such as housing and protected horticulture.

Dr. Lakhmi C. Jain Ph.D., M.E., B.E. (Hons), Fellow (Engineers Australia), is with the University of Piraeus, Athens, Greece. Professor Jain serves the KES International for providing a professional community the opportunities for publications, knowledge exchange, cooperation, and teaming. Involving around 5,000 researchers drawn from universities and companies world-wide, KES facilitates international cooperation and generate synergy in teaching and research. KES regularly provides networking opportunities for professional community through one of the largest conferences of its kind in the areas of KES.

Part I
Main Track

Chapter 1
Early Adopters Versus Late Adopters: Estimating Accuracy of Concept Testing by Selecting Respondents Based on Diffusion of Innovations

Takumi Kato, Yu Zhu, Yusuke Nagata, Junnosuke Kubo, Tomoya Matsue, Yuta Tanaka, Takahiko Umeyama, and Susumu Kamei

Abstract To improve the quality of concept testing, it is necessary to appropriately design it from the perspectives of who (extraction of respondents), what (information to present), and how (information presentation method). Here, we focused on who factors. Since concept testing requires survey respondents to understand the target concept, existing literature has advocated hiring people familiar with the target product. However, when developing new products, the concept should be innovative. There is little need for concept testing for products already familiar to consumers. Therefore, this study introduced innovation diffusion in selecting respondents for concept testing. We distinguished between early adopters (innovators, early adopters, and early majority) and late adopters (late majority and laggards). We derived the following hypotheses: H1: Estimating market share based on concept testing responses is more accurate for early adopters than late adopters. H2: The relationship between concept testing responses and actual purchase behavior is stronger for early adopters than for late adopters. We conducted concept testing on washing machines (drum and vertical types) in Japan online with 1,200 people and compared the results with actual sales data. Therefore, H1 and H2 were supported. This method, identifying appropriate respondents with a single question, would be useful for businesses as it does not increase survey costs or labor for respondents.

T. Kato (✉)
Meiji University, Tokyo, Japan
e-mail: takumi_kato@meiji.ac.jp

Y. Zhu · Y. Nagata · J. Kubo · T. Matsue · Y. Tanaka · T. Umeyama · S. Kamei
Cross Marketing Inc, Tokyo, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_1

1.1 Introduction

Concept testing aims to estimate customer demand before resources are committed to developing a physical prototype [1]. If consumers are not attracted to the product brand concept, it can save investments in development projects [2]. It is particularly important in durable consumer goods, where much capital is required for product development [3]. Thus, concept testing builds a company's innovation capabilities [4].

Improving the quality of concept testing is important, but concept testing is highly confidential in companies [5]. Practitioners prefer to keep their knowledge proprietary, so even though much concept testing is carried out annually, there is little sharing of desirable designs [6]. Therefore, compared to the amount of concept testing performed in daily business, there is surprisingly little academic literature addressing quality factors [7].

Existing literature on the quality of concept testing can be categorized into who, what, and how. Here, we focused on who factors. This study focused on the Japanese washing machine market. It clarified the difference in the accuracy of market share estimation in concept tests between early adopters and late adopters based on the diffusion of innovations. To our knowledge, this study is the first to quantitatively clarify differences in market share accuracy by subject condition in concept testing. These results expand the academic knowledge of concept testing and provide meaningful, practical suggestions.

1.2 Literature Review and Hypothesis Development

1.2.1 Information to Present (What)

Attractive concepts are a source of consumer loyalty [8]. However, the concepts of market products and services are often abstract, and it may be necessary to present information that can be visualized concretely. The following three types of information are commonly used. The first is product design, which is a source of product competitiveness and a significant driver of purchasing behavior [9]. Second, product and corporate brands influence consumer perception. For unknown product brands, the sound of the name and its consistency with product features also influence consumer perception [10]. Although the product brand is equally vital, the corporate brand has a greater impact on the concept test. This is because, even with an excellent product concept, consumer reactions differ depending on the company that provides it [11]. Third, price is used by consumers to judge quality [12].

1.2.2 Information Presentation Method (How)

Benefiting from the development of 3D printers and virtual reality (VR), presenting products through rapid prototyping is possible. This method is currently widely used in new product development [13]. In particular, VR, which does not require physical manufacturing or a large physical space, is effective for considering consumer needs in the early stages of development. Concept testing is also an effective means of communicating product concepts to consumers by combining all the advantages of text and visuals [14]. However, if the target concept is imagined, the change in the ability to predict purchase intention provided by the prototype is minimal [15].

1.2.3 Extraction of Respondents (Who)

Concept testing requires survey respondents to understand the concept in question. Accordingly, existing literature has advocated hiring people familiar with the target product. For example, people with past purchasing experience are more likely to relate their responses to concept testing to their subsequent behavior [16]. For popular products, general consumers have high response accuracy, but for unfamiliar products, experts should be employed [17]. However, the concept should be innovative in new product development. Otherwise, it will simply be an improvement on an existing product. Thus, there is little need for concept testing for familiar products. Even if the product concept is innovative, many consumers can understand the content well [17]. This study introduced innovations [18] in selecting respondents for concept testing. It defined two categories: early adopters (innovators, early adopters, and early majority) and late adopters (late majority and laggards). Then, we derived the following hypotheses.

H1: Estimating market share based on concept testing responses is more accurate for early adopters than late adopters.
H2: The relationship between concept testing responses and actual purchase behavior is stronger for early adopters than for late adopters.
H3: Early adopters spend more time viewing concept information during concept testing than late adopters.

1.3 Method

An online survey was conducted from November 13 to 17, 2023, targeting washing machines (drum and vertical types) in Japan. The survey was distributed through a survey panel held by Cross Marketing Inc. We collected surveys so that each attribute was evenly distributed to avoid biasing the data toward any particular attribute. Specifically, the survey was distributed completely randomly to the panel. All study

participants provided informed consent. The sample size was 600 people for each type, totaling 1,200 people. The target brands were the top five in sales for each type: drum type (Sharp, Panasonic, Toshiba, Hitachi, and Aqua) and vertical type (Hitachi, Panasonic, Toshiba, Haier, and Sharp). Point-of-sales data from January to December 2022 were obtained from GfK Japan. In addition to gender and age, the attribute items obtained in the survey included value for product adoption based on innovations [18]. After that, five concept sheets, as shown in Fig. 1.1, were presented, and the participants were asked to name the brand they would most like to purchase. Finally, participants were asked about their purchase probability using a 7-point Likert scale.

To verify H1, we first conducted a chi-square test periodically on the value for product adoption matrix (early adopter and late adopter) × brands respondents most want to purchase, to confirm differences in distribution. Then, we evaluated the error with sales results. To verify H2, we defined people who selected the top two choices based on the 7-point Likert scale as "people with actual purchase intention," differences in the proportions were evaluated using a chi-square test. In H3, we measured the total viewing time of the five concept sheets and verified the differences using the Brunner–Munzel test.

Fig. 1.1 Example of concept sheet (Toshiba's ZABOON)

1.4 Results and Discussion

1.4.1 Results

Table 1.1 shows that early adopters accounted for 42.5%, and late adopters 57.5%. As a result of applying the chi-square test to the value for product adoption matrix (early adopter and late adopter) × brand, p-value = 0.027 for drum type and p-value < 0.000 for vertical type; significant differences were detected at the 5% level. In other words, the distribution of brands that early and late adopters want to purchase differs. Table 1.2 shows the concept testing and actual share distribution for drum type, and Table 1.3 for vertical type. Figure 1.2 shows the results of evaluating the deviation from the actual share using mean absolute percentage error (MAPE). The results show that the error is about 2.5% lower for drum and vertical types for early adopters, supporting H1.

Table 1.1 Respondent attributes

Item	Breakdown	Drum type		Vertical type	
		Number of respondents	Ratio (%)	Number of respondents	Ratio (%)
Gender	Male	308	51.3	312	52.0
	Female	292	48.7	288	48.0
Age	20 s	148	24.7	133	22.2
	30 s	146	24.3	118	19.7
	40 s	114	19.0	126	21.0
	50 s	106	17.7	113	18.8
	60 s	86	14.3	110	18.3
Values for product adoption	Late adopter	255	42.5	340	56.7
	Early adopter	345	57.5	260	43.3

Table 1.2 Share of concept test respondents and actual sales of washing machines (drum type)

Brand	Late adopter		Early adopter		Sales	
	Number of respondents	Share (%)	Number of respondents	Share (%)	Unit	Share (%)
Sharp	90	35.3	91	26.4	57,499	25.0
Panasonic	33	12.9	68	19.7	56,685	24.7
Toshiba	25	9.8	32	9.3	47,837	20.8
Hitachi	57	22.4	98	28.4	43,360	18.9
Aqua	50	19.6	56	16.2	24,281	10.6
Total	255	100.0	345	100.0	229,662	100.0

Table 1.3 Share of concept test respondents and actual sales of washing machines (vertical type)

Brand	Late adopter		Early adopter		Sales	
	Number of respondents	Share (%)	Number of respondents	Share (%)	Unit	Share (%)
Hitachi	82	24.1	56	21.5	136,321	28.8
Panasonic	46	13.5	60	23.1	93,549	19.8
Toshiba	24	7.1	36	13.8	91,848	19.4
Haier	60	17.6	34	13.1	80,737	17.1
Sharp	128	37.6	74	28.5	70,524	14.9
Total	340	100.0	260	100.0	472,979	100.0

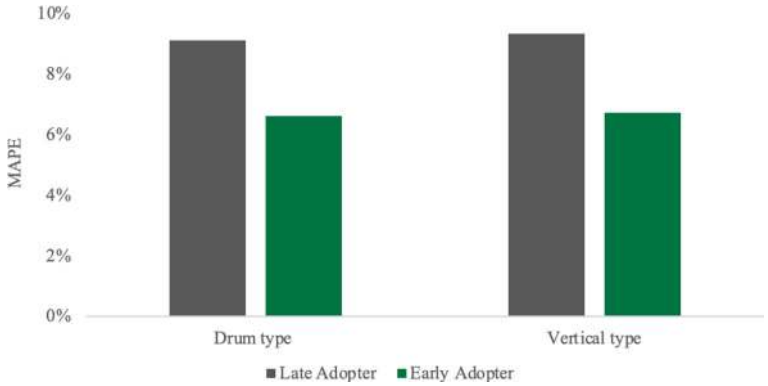
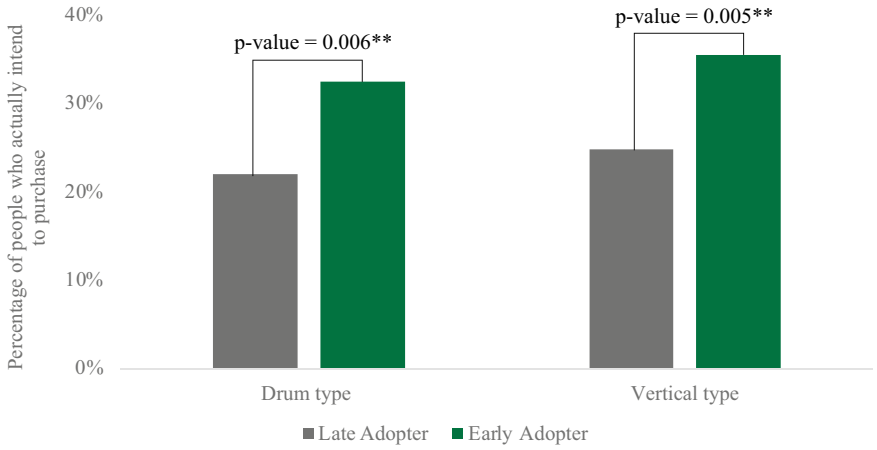


Fig. 1.2 Estimated accuracy of concept test by value for product adoption

Next, Fig. 1.3 shows that the actual purchase intention retention rate is more than 10% higher among early adopters. As a result of the chi-square test, p-value = 0.006 for drum type and p-value = 0.005 for vertical type, and significant differences were detected at the 5% level. Therefore, early adopters could answer assuming actual purchase, supporting H2.

Finally, Figs. 1.4 and 1.5 show the distribution of concept confirmation times. For the drum type, the average viewing time for the five concept sheets was 61.973 s for early adopters and 64.738 s for late adopters. The result of the Brunner–Munzel test was p-value = 0.302, and no significant difference was detected. Similarly, for the vertical type, the time was 63.786 s for early adopters and 65.793 s for late adopters. The result of the Brunner–Munzel test was p-value = 0.228, and no significant difference was detected. Therefore, H3 was not supported. Throughout the test, there are no major differences between the drum and vertical types.



***p<0.001; **p<0.01; *p<0.05.

Fig. 1.3 Difference in the percentage of people who intend to purchase

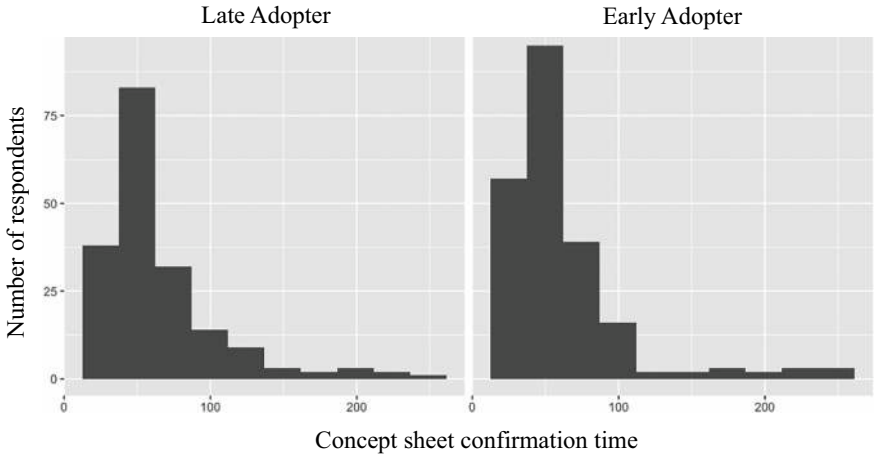


Fig. 1.4 Concept sheet confirmation time for washing machines (drum type)

1.4.2 Practical Implications

This study provides two practical implications. First, concept testing should target consumers with early adopter values based on innovation diffusion. In concept testing, it is difficult to present complete product information, so respondents need imagination to understand the concept. This method, identifying consumers with that ability with a single question, is easy to apply because respondents require less research cost and effort.

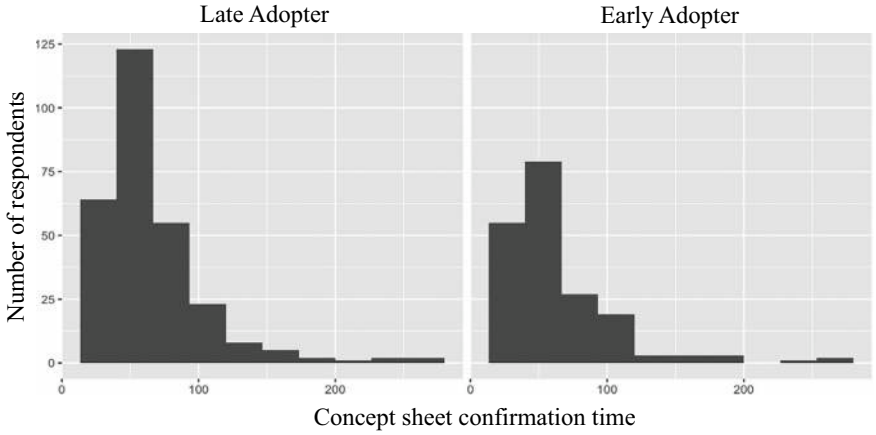


Fig. 1.5 Concept sheet confirmation time for washing machines (vertical type)

Second, concept confirmation time has little effect on answer quality. This would be counterintuitive. In general, it is feared that people who respond quickly will have lower response quality [19]. However, since the time it takes to understand a problem and come up with an answer depends on the person, people with quick thinking can provide appropriate answers even in a short time. Especially when compared to late adopters, early adopters take less time to understand concepts. Therefore, screening should not be based solely on concept viewing time.

1.4.3 Limitations and Future Work

This study has several limitations. First, the results are limited to washing machines in Japan, so product and country expansion is essential. Importantly, it can be applied to durable and consumer goods, the hospitality industry, and IT services. Second, this study tests existing products to confirm the difference in market share. Consequently, verifying even unreleased products that consumers have never seen is desirable. Therefore, it is necessary to link concept testing results before the product's release and the sales share after its release. Third, there is still much room to consider classifying values. This study adopted diffusion of innovations, but estimation accuracy may be further improved by using other concepts. Additionally, this study asked respondents about their values regarding all kinds of products. However, asking about their values only for the target product may be more appropriate. These are future research topics.

1.5 Conclusion

This study proposed a method for selecting respondents based on the diffusion of innovations to improve the quality of concept testing. In other words, it is a factor that corresponds to "who" when improving the quality of concept testing. In addition, it is also necessary to consider "what" and "how" factors and accumulate knowledge about the design of concept testing comprehensively. Although concept testing is carried out daily in business, there is a lack of knowledge in academic research due to the high level of confidentiality, a barrier to information regarding new products. There is also a concern that researchers may assume that it is difficult for consumers to understand and evaluate abstract objects such as concepts. However, when consumers are selected appropriately, they can fully understand the benefits of the concept. It is essential to ensure the validity of the results without relying on the intuition of each person in charge of concept testing design to prevent companies from making unnecessary investments. This major issue is directly linked to management in durable consumer goods such as cars, where new product development requires much time and money. Therefore, researchers should continue accumulating knowledge in this area, and practitioners should establish standard designs based on academic knowledge.

Acknowledgements This work was supported by JSPS KAKENHI (23K12567).

References

1. Crawford, M.C., Di Benedetto, A.C.: New Products Management, 8th edn. McGraw-Hill, New York (2006)
2. Page, A.L., Rosenbaum, H.F.: Developing an effective concept testing program for consumer durables. J. Prod. Innov. Manag.Innov. Manag. **9**(4), 267–277 (1992). https://doi.org/10.1111/1540-5885.940267
3. Acito, F., Hustad, T.P.: Industrial product concept testing. Ind. Mark. Manag.Manag. **10**(3), 157–164 (1981). https://doi.org/10.1016/0019-8501(81)90011-0
4. Durmuşoğlu, S.S., Barczak, G.: The use of information technology tools in new product development phases: analysis of effects on new product innovativeness, quality, and market performance. Ind. Mark. Manag.Manag. **40**(2), 321–330 (2011). https://doi.org/10.1016/j.indmarman.2010.08.009
5. Antikainen, M., Mäkipää, M., Ahonen, M.: Motivating and supporting collaboration in open innovation. Eur. J. Innov. Manag.Innov. Manag. **13**(1), 100–119 (2010). https://doi.org/10.1108/14601061011013258
6. Peng, L., Finn, A.: Concept testing: the state of contemporary practice. Mark. Intell. Plan.Intell. Plan. **26**(6), 649–674 (2008). https://doi.org/10.1108/02634500810902884
7. Peng, L., Finn, A.: How cloudy a crystal ball: a psychometric assessment of concept testing. J. Prod. Innov. Manag.Innov. Manag. **27**(2), 238–252 (2010). https://doi.org/10.1111/j.1540-5885.2010.00712.x
8. Kato, T.: An empirical study of brand concept recall as a predictor of brand loyalty for Dyson. From grand challenges to great solutions: digital transformation in the age of COVID-19. WeB **2021**, 76–86 (2021). https://doi.org/10.1007/978-3-031-04126-6_7

9. Canto Primo, M., Gil-Saura, I., Frasquet-Deltoro, M.: The role of marketing and product design in driving firm's performance. J. Prod. Brand. Manag.Manag. **30**(2), 231–243 (2021). https://doi.org/10.1108/JPBM-07-2019-2477
10. Zamudio, C., Jewell, R.D.: What's in a name? Scent brand names, olfactory imagery, and purchase intention. J. Prod. Brand. Manag.Manag. **30**(2), 281–292 (2020). https://doi.org/10.1108/JPBM-06-2019-2418
11. Kato, T., Kamei, S., Ootsubo, T., Ichiki, Y.: More information is not better: examining appropriate information for estimating sales performance in concept testing. J. Bus. Anal. **6**(3), 188–202 (2023). https://doi.org/10.1080/2573234X.2023.2167670
12. Lee, J.E., Chen-Yu, J.H.: Effects of price discount on consumers' perceptions of savings, quality, and value for apparel products: mediating effect of price discount affect. Fash. Text. **5**, 1–21 (2018). https://doi.org/10.1186/s40691-018-0128-2
13. Tih, S., Wong, K.K., Lynn, G.S., Reilly, R.R.: Prototyping, customer involvement, and speed of information dissemination in new product success. J. Bus. Ind. Mark. **31**(4), 437–448 (2016). https://doi.org/10.1108/JBIM-09-2014-0182
14. Peng, L., Cui, G., Li, C.: Individual differences in consumer responses to traditional versus virtual concept testing. J. Prod. Brand. Manag.Manag. **21**(3), 167–175 (2012). https://doi.org/10.1108/10610421211228784
15. García-Milon, A., Martínez-Ruiz, M.P., Olarte-Pascual, C., Pelegrin-Borondo, J.: Does the product test really make a difference? Evidence from the launch of a new wine. Food Qual. Prefer. **71**, 422–430 (2019). https://doi.org/10.1016/j.foodqual.2018.08.007
16. Ozer, M.: The moderating roles of prior experience and behavioral importance in the predictive validity of new product concept testing. J. Prod. Innov. Manag.Innov. Manag. **28**(1), 109–122 (2011). https://doi.org/10.1111/j.1540-5885.2010.00784.x
17. Heiskanen, E., Hyvönen, K., Niva, M., Pantzar, M., Timonen, P., Varjonen, J.: User involvement in radical innovation: are consumers conservative? Eur. J. Innov. Manag.Innov. Manag. **10**(4), 489–509 (2007). https://doi.org/10.1108/14601060710828790
18. Rogers, E.M.: Diffusion of Innovations, 5th edn. Free Press (2003)
19. Revilla, M., Ochoa, C.: What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? Soc. Sci. Comput. Rev. Comput. Rev. **33**(1), 97–114 (2015). https://doi.org/10.1177/0894439314531214

Chapter 2
Cascading Sum Augmentation: Leveraging Populated Feature Spaces

Cristian Simionescu, Robert Herscovici, and Cosmin Pascaru

Abstract Data augmentation techniques are pivotal in enhancing the generalization capabilities of deep artificial neural networks. Traditional approaches in image augmentation often prioritize generating samples comprehensible to human observers, inadvertently overlooking a spectrum of data potentially beneficial for neural network training. This paper introduces a novel data augmentation technique for image classification tasks, termed **Sum Augmentation**, which expands upon the concept of linear interpolation of inputs. Our method distinctively combines a substantially larger number of data points, substantially expanding the dataset size. We demonstrate the utility of generating up to eight combined samples to produce complex images, which, though seemingly unintelligible, enable deep models to extract valuable insights. We propose **Cascading Sum Augmentation**, a novel training procedure that effectively transfers this knowledge, enhancing model generalization. Our findings indicate a more pronounced accuracy improvement in smaller datasets. We also introduce a derived test-time augmentation technique from **Sum Augmentation** to further boost performance.

2.1 Introduction

The landscape of Deep Learning has seen remarkable advancements, significantly impacting various Computer Vision tasks such as classification, object detection, and image segmentation. Despite these advancements, training Deep Artificial Neural Networks (DANNs) often requires extensive labeled data, which can be costly and challenging to acquire, particularly in specialized fields like robotics or medical imaging. This challenge underlines the importance of innovative data augmentation strategies to enhance model generalization. The heightened relevance of security in

C. Simionescu (✉) · R. Herscovici · C. Pascaru
Alexandru Ioan Cuza University, Iasi, Romania
e-mail: cristian@nexusmedia.ro

C. Pascaru
e-mail: cosmin.pascaru@info.uaic.ro

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_2

modern contexts [1, 2] further emphasizes the need for neural networks with robust generalization abilities, as they are not only more effective but also more resilient to adversarial attacks. Various data augmentation techniques have been developed, each varying in success and applicability across different datasets, models, and machine learning methods. These techniques are integral to applications ranging from image classification [3] to speech recognition [4, 5], predominantly resulting in human-interpretable images through transformations like flipping, cropping, rotation, and brightness adjustments [6–9].

Recent studies in image data augmentation [10–13] highlight the efficacy of techniques beyond humanly perceptible image modifications, noting their positive impact on neural network generalization.

This paper introduces **Sum Augmentation**, a generalized method for generating linear combinations of data points (Sect. 2.3), demonstrating significant error rate reductions on CIFAR-10/100 datasets. Inspired by this technique, we propose **Test-Time Sum Augmentation** (Sect. 2.5), a novel test-time augmentation method to further enhance performance and strengthen model resilience against adversarial attacks.

2.2 Related Work

In the field of data augmentation for imbalanced datasets, seminal work by Chawla et al. [14] introduced the concept of synthesizing new examples by interpolating among nearest neighbors within the minority class. Extending this idea, DeVries et al. [15] explored interpolations within the same class to bolster generalization capabilities.

The principle of linear interpolations underpinning feature vectors and their corresponding targets was methodically examined by the Mixup technique [10], which posits that such interpolations construct a combinatorial manifold within the feature space. This combinatorial approach, being essentially a piecewise linear manifold, complicates the adversarial landscape and simultaneously smoothens the loss landscape, thus promoting superior generalization and expedited convergence within the gradient descent framework. Despite the reported success of Mixup in enhancing generalization across a spectrum of tasks—including image classification and speech recognition—initial investigations by the authors suggested diminishing returns when extending the interpolation beyond pairs of examples, a limitation we revisit and contend in our work.

Furthering the discourse, SamplePairing [13] proposed an additional fine-tuning phase utilizing the unmodified training set. Our findings corroborate the effectiveness of this strategy, particularly its pronounced impact on smaller datasets. Contrary to the claim by Inoue et al. that performance remains stable regardless of whether targets are linearly combined, our results indicate a marked degradation in performance when targets are not appropriately adjusted.

Moreover, Summers et al. [16] presented an alternative strategy that harnesses multiple non-linear combinations to generate new training samples. Their method demonstrated competitive, and in some instances superior, effectiveness compared to both Mixup [10] and Between-Class learning [11].

Our research contributes to this evolving narrative by proposing a generalized training schema that not only amalgamates but also enhances the Mixup, Between-Class, and SamplePairing methodologies to realize additional improvements in test accuracy when integrating more than two inputs. This stands in stark contrast to the prevailing sentiment in prior research, which posited a decline in performance with the inclusion of more samples.

In tandem with these theoretical advancements, we introduce a novel test-time data augmentation strategy specifically optimized to elevate the accuracy of models trained on mixed data samples. This tailored augmentation is designed to leverage the mixed nature of the data to achieve better model performance and robustness, further contributing to the literature on effective training and generalization strategies for deep learning models.

2.3 Sum Augmentation

In this work, we introduce **Sum Augmentation**, a novel approach that generalizes the concept of data augmentation through linear combinations of input data points, building upon the foundations laid by SamplePairing [13] and MixUp [10]. These methods traditionally generate a new augmented data sample x_{new} using the formula:

$$x_{new} = \lambda * x_i + (1 - \lambda) * x_j, \quad x_{i,j} \in D, \text{ where } D \text{ is the original dataset.}$$

As highlighted by Inoue [13], a significant performance enhancement was observed when $\lambda = 0.5$, a finding that aligns with the results from our experiments. To extend this concept for integrating an arbitrary number of samples K, facilitating a broader linear combination, we propose

$$x_{new} = \frac{1}{K} \sum_{i=1}^{K} x_i, \quad x_i \in D, \text{ the original dataset.}$$

This extension not only broadens the potential input domain but also significantly increases it, with the new domain size estimated by

$$|D_{new}| = \binom{|D|}{K}.$$

It is important to note that this is an approximation, as there could be instances where $\frac{1}{K}\sum_{i=1}^{K} x_i$ might already exist within D, though such overlaps are anticipated to be rare in practical image datasets.

For efficient batch processing, our method constructs an augmented batch $batch_{new}$ from a given batch of m samples as follows:

$$batch_{new}[i] = \frac{1}{K}\sum_{j=0}^{K-1} batch[i + j * \left\lfloor \frac{m}{K} \right\rfloor], \quad i \in \left[1, \left\lfloor \frac{m}{K} \right\rfloor\right].$$

This procedure divides the original batch into K sum groups of equal size, with each element in $batch_{new}$ calculated as the average of the corresponding elements across all groups.

Furthermore, we extend this linear combination to the one-hot encoded labels, setting the neural network's target for each augmented sample as

$$target[i] = \frac{1}{K} * c_i,$$

where c_i denotes the count of samples with label i in the sum group. The loss is then computed using binary cross entropy, scaled down by a factor of K.

Visual examples of images generated through Sum Augmentation for different values of K are illustrated in Figs. 2.1, 2.2, and 2.3. While these images retain structural elements, identifying specific objects becomes progressively more challenging with increasing K values, underscoring the complexity of the augmentation process.

Fig. 2.1 $K = 2$ Generated images

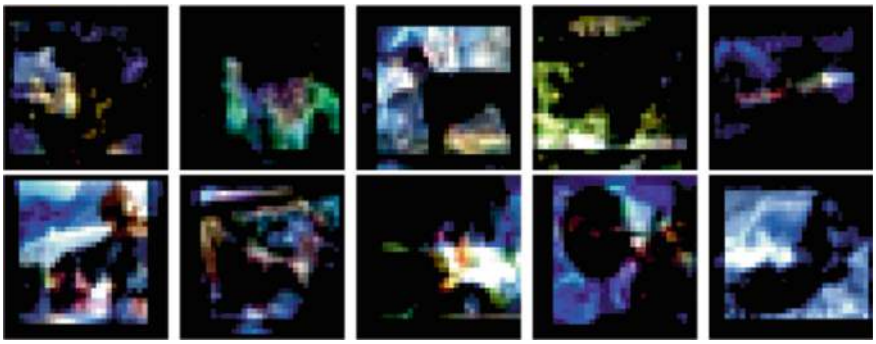


Fig. 2.2 $K = 4$ Generated images

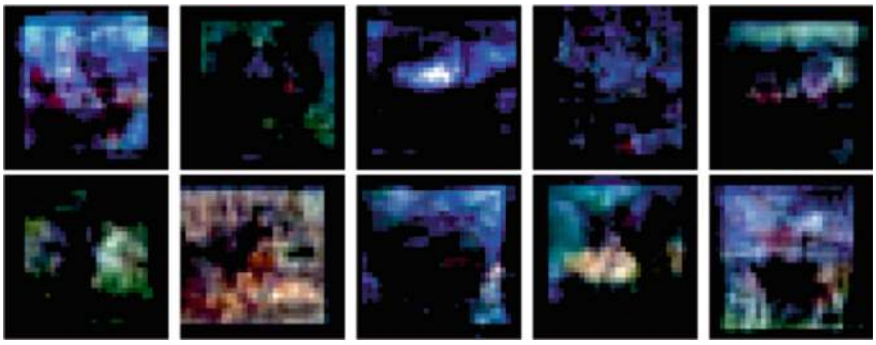


Fig. 2.3 $K = 8$ Generated images

2.4 Cascading Sum Augmentation

The **Sum Augmentation** technique is versatile, accommodating any number of sum groups. We exploit this flexibility through a "cascading" training approach. Initially, the model is trained using K sum groups. A checkpoint is saved each time a new best test accuracy is achieved. When performance plateaus, we transfer the weights from the best checkpoint and continue training with $K/2$ sum groups. This process repeats until $K = 1$, culminating in a final fine-tuning phase using the original training dataset.

1: **procedure** CASCADESUMGROUPS(K)
2: Load the previous model checkpoint
3: **if** $K < 1$ **then return**
4: **else**
5: **while** loss is decreasing **do**
6: Update model parameters
7: **end whilereturn** $CascadeSumGroups(\lfloor \frac{K}{2} \rfloor)$
8: **end if**
9: **end procedure**

2.5 Test-Time Sum Augmentation

Deep neural networks often develop non-robust features [17], crucial for performance yet vulnerable to adversarial attacks [18]. These models may overfit, misclassifying images containing small, feature-specific segments. This vulnerability is exploited in adversarial attacks, such as the One Pixel Attack [19].

Expanding on **Sum Augmentation**, we propose a novel test-time augmentation method to enhance the accuracy of models trained with $K = 2$ sum augmentation and mitigate adversarial attacks. Given an image B_i with target $Target_i$, this method involves the following steps with parameter C:

1. Generate C "augmented" images, each being a linear combination of the original image and a random image from the test dataset. $RandomTestSample()$ returns a random test image:

$$Aug_i = \lambda B_i + (1 - \lambda) RandomTestSample(), \quad \lambda \in [0, 1].$$

   We set $\lambda = 0.5$ for the remainder of this article.
2. Generate predictions for each augmented image using the model, resulting in C 1D tensors (each of size $num\_classes$):

$$Pred_i = predict(Aug_i).$$

3. Compute the mean of all predictions to determine the output of the network:

$$Out = (Pred_1 + Pred_2 + ... + Pred_C)/C.$$

4. The output, Out, is a 1D tensor of size $num\_classes$. The $argmax$ of Out is compared against $Target_i$ to compute performance metrics such as accuracy, loss, and F-score.

This method aims to mitigate the impact of over-fitted features in classification. By averaging predictions over multiple augmented images, the method demands adversaries to craft noise effective across all augmented samples, thus elevating the model's resilience against adversarial attacks.

Preliminary experiments indicate enhanced robustness against adversarial attacks with this method. We hypothesize that generating effective noise for all augmented images is considerably challenging, thereby requiring adversaries to consistently produce classification errors across the majority of augmented data points.

2.6 Experimental Evaluation

To demonstrate the efficacy of our generalized approach, we employed the CIFAR-10 and CIFAR-100 datasets. Data augmentation techniques serve to effectively enlarge the dataset at our disposal, yielding substantial performance enhancements, particularly with smaller datasets. To emulate conditions of limited data availability, we generated subsets of our datasets in various sizes, which facilitated an evaluation of the performance gains potentially afforded by our methodology. Such insights are particularly pertinent for domains like medical imaging, where procuring additional data can be prohibitively expensive or impractical. We utilized subsets of the following sizes for our analysis:

- CIFAR-10
  - 1000 samples (100 samples per class)
  - 5000 samples (500 samples per class)
  - 50000 samples (the original dataset)
- CIFAR-100
  - 10000 samples (100 samples per class)
  - 50000 samples (the original dataset)

The examination of both CIFAR datasets is crucial to substantiate that an increase in the number of classes does not compromise the performance enhancements delivered by our method. In our pursuit to ascertain whether our methodology sustains performance improvements with increased model scale, we selected two distinct WideResNet architectures [20] for training:

- A model with 40 layers and a widening factor of 4, comprising 8.9 million parameters.
- A model with 28 layers and a widening factor of 10, encompassing 36.5 million parameters.

We standardized training across all models with a batch size of 100 and an initial learning rate of 0.05. The learning rate was halved upon observing a plateau in learning for 300 gradient update steps. Stochastic Gradient Descent [21], complemented by Nesterov accelerated momentum [22] of 0.9, was employed as the optimizer, with an $L2$ regularization penalty set to 0.0001. Uniform parameters across all experiments ensure a fair comparison. It is plausible that an exhaustive search for optimal meta-parameters could further elevate the peak performance of **Sum Augmentation** beyond the results presented herein.

2.6.1 Cascading Sum Augmentation Results

A salient observation is that employing **Cascading Sum Augmentation** with larger values of K results in a reduced $L2$ norm of the final model parameters (Fig. 2.4), underscoring the regularization benefits of **Sum Augmentation** noted in prior studies. This effect likely arises from the ability of models trained with cascading techniques to undergo additional epochs before convergence. Notably, there are distinct increases in the norm values when transitioning from K to $K/2$ sum groups, suggesting that performance could be enhanced by implementing more gradual cascading steps.

An intriguing consideration is whether models with high-order Sum Groups (e.g., $K = 8$) can extract valuable features from seemingly "noisy" averaged images. Contrary to initial assumptions, the accuracy obtained was significantly better than anticipated. For instance, on CIFAR-10 with the WideResNet(28,10) model, Sum Augmentation at $K = 8$ achieved an accuracy of 49.05%, trained solely on augmented images (Fig. 2.3).

Our **Cascading Sum Augmentation** method surpassed the performance of models trained exclusively with binary groups and subsequently fine-tuned on the original dataset, as suggested in the literature [10, 13], across all tested scenarios by a notable margin.

As illustrated in Table 2.1, in data-scarce environments, the model with greater capacity, WideResNet(28,10), attained the lowest error rates. This was particularly evident when commencing the Cascading process with $K = 8$ Sum Groups on a CIFAR-100 subset comprising only 100 samples per class. This suggests that higher capacity models can discern useful patterns even when trained on aggregated image inputs. For larger datasets or smaller models, a $K = 4$ configuration consistently yielded superior results. We designate the standard training on the original dataset without augmentation as our "Baseline" against which we gauge the accuracy improvements.

In order to compare our method with similar methods, we ran the experiments described in Table 2.1 using the Mixup [10] and SamplePairing [13] methods. In

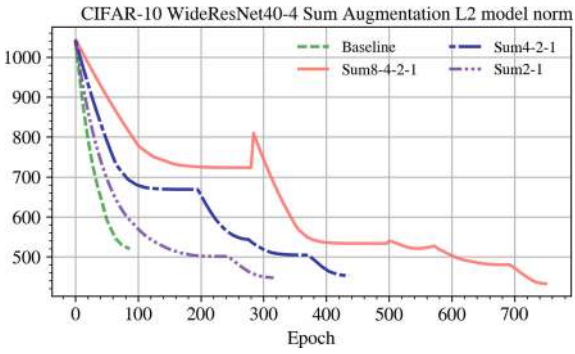


Fig. 2.4 $L2$ norm

Table 2.1 CIFAR-10 and CIFAR-100 test error using cascading sum augmentation. This table compares the test error rates for different Starting Sum Groups across two datasets (CIFAR-10 and CIFAR-100) and configurations of WideResNet (WRN)

SSG	CIFAR-10				CIFAR-100			
	WRN(40,4)		WRN(28,10)		WRN(40,4)		WRN(28,10)	
	100 (%)	500 (%)	100 (%)	500 (%)	100 (%)	500 (%)	100 (%)	500 (%)
8	31.19	13.51	**29.34**	13.21	35.3	20.34	35.01	18.63
4	**30.65**	**13.29**	30.36	**12.81**	**34.7**	**19.85**	**33.01**	**18.09**
2	32.44	14.88	30.99	13.1	35.56	20.17	34.44	18.17
Baseline	47.36	20.15	43.47	19.96	43.2	23.87	41.72	21.94

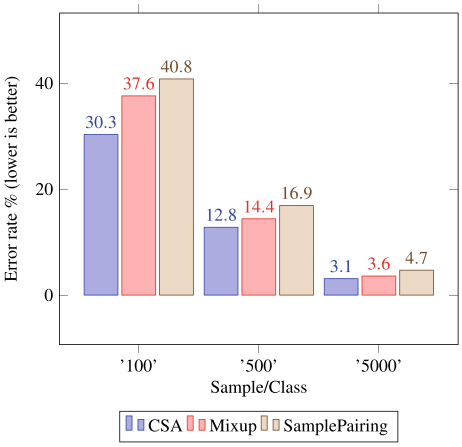


Fig. 2.5 Comparison on CIFAR-10

Fig. 2.5, you can see the comparison between our method CSA (Cascading Sum Augmentation) with $K = 4$ (the most consistent parameter value for the given task) on CIFAR-10 using WideResNet(28,10). Our method's proficiency on small datasets is even clearer when comparing the accuracy boosts each method brings over the baseline performance. Our method brings a 30.15% error reduction while Mixup, the second best performing method, only decreases the error rate by 13.5%. This gap is smaller but still considerable when training on the entire dataset, our method producing a 33.54% decrease in error while Mixup achieves only a 22.86% reduction.

2.6.2 Test-Time Sum Augmentation Results

The results of the experiments using **Test-Time Sum Augmentation** can be found in Table 2.2. These results were obtained by using our test-time data augmentation

Table 2.2 CIFAR-10 and CIFAR-100 test error using test-time sum augmentation. NT: Normal test, TTSA: Test-time sum augmentation. The table presents error rates for different K values under both NT and TTSA for WideResNet(40,4) and WideResNet(28,10) configurations, comparing performance on CIFAR-10 and CIFAR-100 datasets

K **values**	CIFAR-10				CIFAR-100			
	WideResNet(40,4)		WideResNet(28,10)		WideResNet(40,4)		WideResNet(28,10)	
	NT (%)	TTSA (%)	NT (%)	TTSA (%)	NT (%)	TTSA (%)	NT (%)	TTSA (%)
8-4-2	6.53	6.32	5.58	5.11	27.18	24.06	25.11	**20.25**
4-2	6.91	6.32	5.35	**5.09**	27.22	23.9	24.42	20.46
2	6.19	**6.03**	5.74	5.39	27.19	**23.58**	24.32	20.29

method on the best performing weights of models trained with the cascading algorithm with one small alteration. For this method, we chose to stop the cascading algorithm when reaching $K = 2$. The reason for this choice is that all runs that reached $K = 1$ would have worse performance when inferring on data using **Test-Time Sum Augmentation**. This is to be expected since when a model would fine-tune on the normal dataset it would have to adjust to the new sample distribution hence leading to some degree of *Catastrophic forgetting* [23].

As it can be seen from the test error rates, our test-time data augmentation method leads to considerable boosts in prediction performance for models whose final fine-turning is done on **Sum Augmentation** with $K = 2$. Comparing these results with the ones in Table 2.1 it is clear that the method is not suitable for being used when we only want to obtain a model with the best test data accuracy. We consider the use-case where **Test-Time Sum Augmentation** should be applied is when we are either required or desire to utilize models trained on linear combinations of samples. As stated in Sect. 2.5, from limited preliminary experiments our proposed test-time data augmentation was shown capable of improving the robustness of a deep model to adversarial examples.

2.7 Conclusion

We introduced **Cascading Sum Augmentation**, a data augmentation strategy demonstrating significant error rate reductions on CIFAR datasets. The potential for further improvements includes smoother transitions during cascading and exploration of varied sample weights.

Future work should explore the applicability of our method across different datasets and tasks, as well as the intrinsic robustness of models trained with **Cascading Sum Augmentation** against adversarial attacks. Additionally, the behavior of our approach on unbalanced data needs to be explored, as there are reasons to believe it should work better than standard augmentation methods since it allows for greater

levels of combinations; however, the risk of having the manifold space become too sparse due to the smaller classes is also present. **Test-Time Sum Augmentation** offers another promising avenue for investigation, particularly in balancing performance and robustness through the parameter λ.

References

1. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. Mach. Learn. **81**(2), 121–148 (2010)
2. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., Tygar, J.D.: Can machine learning be secure? In: Proceedings of the 2006 ACM symposium on information, computer and communications security, ASIACCS'06, pp. 16–25, New York, NY, USA, 2006. ACM (Nov 2006)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (May 2017)
4. Graves, A., Mohamed, A.-R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE (2013)
5. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: end-to-end speech recognition in English and Mandarin. In: International Conference on Machine Learning, pp. 173–182 (2016)
6. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: training imagenet in 1 hour (2017)
7. LeCun, Yann, Bottou, Léon., Bengio, Yoshua, Haffner, Patrick, et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
9. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation (2017). arXiv:1708.04896
10. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization (2017). arXiv:1710.09412
11. Tokozume, Y., Ushiku, Y., Harada, T.: Between-class learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5486–5494 (2018)
12. Tokozume, Y., Ushiku, Y., Harada, T.: Learning from between-class examples for deep sound recognition (2017). arXiv:1711.10282
13. Inoue, H.: Data augmentation by pairing samples for images classification (2018). arXiv:1801.02929
14. Chawla, N.V., Bowyer, K.W., Hall, L.O., Philip Kegelmeyer, W.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
15. DeVries, T., Taylor, G.W.: Dataset augmentation in feature space (2017). arXiv:1702.05538
16. Summers, C., Dinneen, M.J.: Improved mixed-example data augmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1262–1270. IEEE (2019)
17. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features (2019). arXiv:1905.02175
18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2013). arXiv:1312.6199
19. Su, J., Vargas, D.V., Sakurai, K: One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. (2019)
20. Zagoruyko, S., Komodakis, N.: Wide residual networks (2016). arXiv:1605.07146
21. Ruder, S.: An overview of gradient descent optimization algorithms (2016). arXiv:1609.04747

22. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proceedings of Machine Learning Research, vol. 28, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR (2013)
23. Pfülb, B., Gepperth, A., Abdullah, S., Kilian, A.: Catastrophic Forgetting: Still a Problem for dnns. Lecture Notes in Computer Science, pp. 487–497 (2018)

Chapter 3
Failure Prediction for Large Anti-drone System Clusters

Buket Kazma and Fatih Semiz

Abstract In this study, we aimed to predict the failure of a system that includes multiple anti-drone systems, namely radar and jammer devices. It is anticipated that predictive maintenance, which has been attempted in many areas before, in the anti-drone field will reduce the costs. In this study, we developed a feature selection that suits the requirements of the problem. Afterward, Support Vector Machines (SVM), and Multi-Layer Perceptron Neural Network (MLPNN) models are adapted to our problem. Created models are fed with one hand-crafted single device data set, one synthetic single device data set, and one synthetic multi-device data set and reported the success rates. It is observed that the results provide successful predictions that could used in the field.

3.1 Introduction

Anti-drone systems refer to systems that have civil and military applications aiming to defend certain areas against drones. A typical anti-drone system consists of a command and control center as well as devices that perform detection, identification, and neutralization stages [15]. Radars [20], thermal cameras, RF scanner technologies, optical cameras, acoustic signals emitted from motors, and hybrid systems can be used for detection [7]. Methods such as radio frequency identification (RFID), drone tracking, and flight estimation can be used for identification [22]. Drone hijacking, drone spoofing, geofencing, killer drones, drone capturing, and drone jamming methods can be applied for drone neutralization. In this study, we worked on a system where multiple anti-drone systems, including radars and jammers, are combined and monitored [1, 13, 14].

In such extensive systems, maintenance operators must visit the devices and perform maintenance routinely or when a malfunction occurs [24]. However, mainte-

B. Kazma · F. Semiz (✉)
Aselsan Inc., Ankara, Turkey
e-mail: fatihsemiz@aselsan.com.tr

B. Kazma
e-mail: bkazma@aselsan.com.tr
© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_3

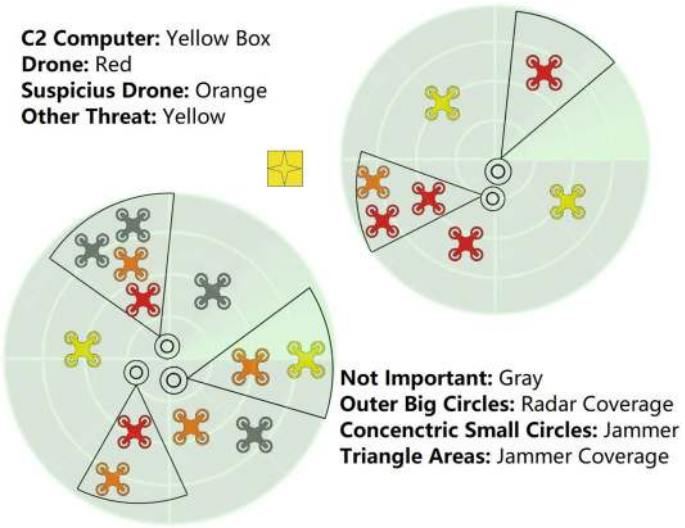


Fig. 3.1 An example problem scenario that includes 2 radar coverage areas, shows the threats detected in those areas in different colors according to their types, and neutralizes them with 5 jammers

nance procedures are typically carried out when malfunctions occur. That leads some devices to be inaccessible in case of instant threat. If we adopt a prediction and monitoring process (Predictive Health Monitoring), the maintenance expenditure will decrease. Besides, more devices will be available in case of emergency threats.

An example scenario of the problem is presented in Fig. 3.1. The yellow star between the two outer big circles represents the command and control (C2) center's location. Each outer big circle shows the radar coverage area. The locations of threats for drone detection and identification within the radar coverage area are provided. The ones detected as drones are shown in red, suspicious drones are shown in orange, other threat types are shown in yellow, and not-important detections are shown in gray. Jammers are represented with concentric small circles and their coverage areas are shown with triangles. In such a problem scenario, it will be important to predict when the devices in the field may need maintenance and keep all devices working before a threat occurs.

Our goal in this work was to apply failure prediction techniques to this field [17] to discover a solution to this problem. Although failure prediction is a popular study in electronics [3], transportation systems [18], aviation [9], etc., there is only a limited number of studies in the military domain. Another challenge is that these studies' datasets were not publicly open. Therefore, we aimed to produce three publicly accessible datasets [8] to encourage researchers in the military domain.

The paper is organized as follows: Section 3.2 construes related work and background. Section 3.3 provides the problem description. Section 3.4 depicts the pro-

posed method. Section 3.5 describes the data generation and experiment results of the proposed method. Section 3.6 concludes the paper and points out future work.

3.2 Related Work and Background

In recent years, with the easier production and acquisition of drones, there has been a significant increase in the number of drone users. Yaacoub et al. [23] provided a survey about drone systems and in that article they stated that there are three types of drones. These 3 types are multi-rotor drones, fixed-wing drones, and hybrid-wing drones. Multi-rotor drones can do vertical take-off and landing. On the other hand, fixed-wing drones have the advantage of being more energy efficient due to their ability to glide and travel at a high speed. Whether a radar can detect a drone or not mainly depends on the frequency and the rotation speed of the radar. The advances and usage increase of drones revealed the need to defend lands against drone threats, particularly in urban areas [16]. Device maintenance becomes a challenge for large anti-drone systems, especially when there are several of them in the area.

In the literature, there are various studies on analyses, including automatically finding and predicting machine errors and understanding when they will fail. Remaining usable life (RUL) lowers the cost of unscheduled maintenance by predicting the time left until a fault with the machinery arises. Lei et al. [11] provided a model-based approach for estimating the RUL of machinery.

Zonta et al. [26] provided a survey about predictive maintenance in the industry. This study states that the main focus of Industry 4.0 is on addressing data analytics and machine learning methods for predictive maintenance problems. The most common prediction approaches are physical-model-based [21], knowledge based [2], and data-driven [6].

Lei et al. [12] published a study that reviews empirical mode decomposition approaches in fault diagnosis of rotating machinery. Rotating machinery is a mechanical equipment that is frequently used in the industry. Maintenance of it is critical because it operates under harsh conditions. For this reason, most of the fault diagnosis studies have focused on rotating machinery applications. This study discusses the applications of empirical mode decomposition, one of the significant signal processing methods, to these issues. Wei et al. [19] provided a support vector domain description (SVDD) to make fault predictions on radar equipment. Furthermore, Khalil et al. [10] in 2020 provided a machine learning-based method to make hardware fault prediction, especially on circuits.

As this is one of the first attempts at using fault prediction in anti-drone systems, we supplied our features and datasets and utilized commonly known techniques in this new area.

3.3 Problem Description

In this problem, the goal is to predict whether the devices whose status is being monitored will enter an error state. An instance of the problem consists of N jammers and M radars. The state of all devices in the system is recorded in every t time interval. For each device; battery status, battery level, heat status, connection status, bit status, and system status are recorded. These constitute BIT (Built-in Test) data of devices. They construct our datasets with other device information: device name, device latitude, device longitude, city name, broadcast on date, and broadcast off date of the device. Details of the features are shown in Table 3.1. Each case should be mapped to 0 representing a failure and the case should be mapped to 1 if there is no failure predicted for that device.

Table 3.1 Details of the features used in our datasets

Feature	Values	Definition
Device name	Any string	Name of the device
Device type	Radar / Jammer	Type of the device
Latitude	(–85.051112878)–(85.051112878)	Latitude of the device
Longitude	(–85.051112878)–(85.051112878)	Longitude of the device
City name	any string	City of the device
Broadcast on date	in datetime64[ns] format	Timestamp of broadcast on
Broadcast off date	in datetime64[ns] format	Timestamp of broadcast off
Bit status	0/1	0: bit report is not prepared 1: bit report prepared
Battery status	0 / 1	0: low battery 1: enough battery
Battery level	[1,5]	Battery level: 1 empty, 5 full
Heat status	0/1	0: no error; 1: error (radar) 0: error; 1: no error (jammer)
Connection status	0/1	0: no error; 1: error
System Status	0/1	0: no error; 1: error

3.4 Method

In this section, we suggest an outline of the failure prediction steps of the devices. The outline consists of two stages: feature selection and model building. The feature selection stage utilized domain-specific knowledge and the Principle Components Analysis (PCA) method. Model-building stage adopted Support Vector Machines (SVM) and Multi-Layer Perceptron Neural Network (MLPNN) models.

3.4.1 Feature Selection

In the feature selection stage, first, "working hours" and "down hours" are derived. Working hours represent the total work hours of the device by finding the difference between when the device was turned off (broadcast off date) and when it turned on (broadcast on date) in that record. Down hours represent the total closed hours of the device by finding the difference between when the device was turned on in that record (broadcast on date) and when the device was last turned off (broadcast off date) from the previous records of the same device. Next, these broadcasting dates are dropped. Lastly, PCA is implemented.

3.4.2 Model Building

In this stage, we adopted support vector machines, and multi-layer perceptron neural network for our datasets.

Support vector machines (SVMs) were introduced by Vapnik et al. [4]. It is a supervised learning model that applies to classification and regression problems and is a member of linear classifiers. The objective is to generate a $n - 1$-dimensional hyperplane that divides an n-dimensional data space in a way that maximizes the margin between two classes.

Hand-crafted single device dataset (HSD) and synthetic single device dataset (SSD) adopted an SVM model using a linear kernel and a C parameter (penalty parameter of the error term) with the value of 1. The model uses the first 250 instances for training and the last 50 instances for testing in both datasets. In that case, the train-test split of the datasets is approximately 86% train and 14% test.

Multi-layer perceptron neural network (MLPNN) is a name for a modern feed-forward artificial neural network (ANN) consisting of fully connected neurons with a nonlinear kind of activation function, organized in at least three layers, notable for being able to distinguish data that is not linearly separable [5].

Synthetic multiple device dataset (SMD) adopted an MLPNN model. Before using an MLPNN, data is fit to a min-max scaler. It is a typical method for scaling features between the range [0,1] or [−1, 1]. During the training, Adam optimizer and binary

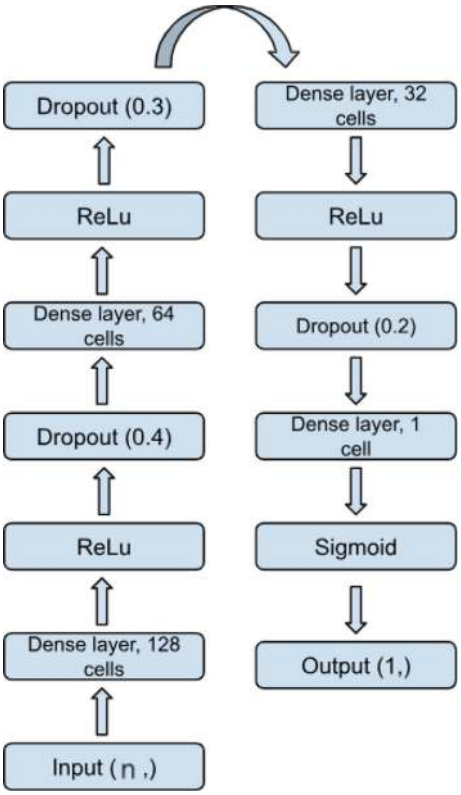


Fig. 3.2 The stages of our multi-layer network used in the study are summarized in the picture above. The model first starts by taking a $n\_components \times 1$ feature set as input and then converts it to a 1×1 prediction value through various stages

cross-entropy loss were chosen. The model is trained in 15 epochs using a batch size of 64. The model uses randomly selected 85 instances for training and randomly selected 15 instances for testing. While training, the validation split is chosen as 0.12. In that case, the train-validation-test split of the dataset is 74.8% train, 10.2% validation, and 15% test. Details of the MLPNN model are in the Fig. 3.2.

3.5 Experimental Study

In this section, the prediction performance of the machine learning models defined in Sect. 3.4 is evaluated. Since there was no previous study on maintenance in this field, we created our own data sets. The following sections explain the data sets created. Afterward, the results of the developed algorithms are compared.

3.5.1 Data Generation

The rationale behind our data creation is utilizing the most common data of radar and jammer devices in the domain. That is useful particularly when sensor data (vibration, dust level, etc.) is unavailable in the device environment. Battery status, battery level, heat status, connection status, bit status, and system status constitute BIT (Built-in Test) data. The BIT is an important technology that can greatly improve the testability and diagnosis capability of the system [25]. To conclude, failure situations can be predicted from our datasets. This can be used for PHM purposes.

3.5.1.1 Hand-Crafted Single Device Dataset

At first, we wanted to create a small-scale hand-crafted dataset consisting of 300 instances. With that, some attributes of the dataset to comply with real-life situations could be structured. The dataset demonstrates only one type of device radar.

Device name, device type, latitude, longitude, city name, and bit status are constant values. Heat status and connection status are randomly determined. Broadcast on date, and broadcast off date are random between available periods. The available periods apply these rules:

- Broadcast on a date varies between +6 to 26 hours from the broadcast off date of the previous instance
- Broadcast off date varies between +0 to 48 hours from broadcast on the date of that instance.

Other features are the ones we have hand-crafted within a real-life logic. Battery level is decreasing from 5 to 0 by time in the records. Battery status is compatible with the battery level. For example, the battery status is 1 (enough battery) when the battery level is 4. System status is 1 when the connection status is 1 or the battery level is low and the heat level is high. Otherwise, it is more likely to be 0 but there are opposite situations too. Lastly, the label is the "system status" in this dataset.

3.5.1.2 Synthetic Single Device Dataset

After generating a small-scale hand-crafted dataset, we wanted to generate it synthetically. This time, instead of determining heat status manually, we filled it out with code applying a rule. If the working hours are more than 20 hours or vary between 15 hours and 2*down hours, the heat status is 1; otherwise, it is 0 . Connection status is again a random value. System status (label of the prediction) is filled out with code as shown in Table 3.2. All other attributes remain the same from the previous dataset. Since this dataset is a half-auto-generated version of the previous one, it has 300 instances too. Lastly, the label is "failure" in this dataset.

Table 3.2 Possible system status values according to heat and connection status

Heat status	Connection status	System status
0	0	0
1	1	1
0	1	1
1	0	0/1

3.5.1.3 Synthetic Multiple Device Dataset

Finally, we developed an algorithm that takes the sample size as an input and auto-generates different types of devices that conform to the rules. Broadcast on date, broadcast off date, and bit status abide by previous rules. Devices are randomly selected from the Table 3.3. The rest of the features are auto-generated by the algorithm within all the values they can take in real-life scenarios. Failure (prediction label) is also auto-generated. It just considers how many percent it is likely to be 0 or 1 in that specific case regarding bit status. These cases are shown in Table 3.4.

Table 3.3 Available device's information table

Device Id	Device name	Device type	Latitude	Longitude	City name
1	Device 1	Radar	38660	40400	Diyarbakir
2	Device 1	Radar	38600	40396	Diyarbakir
3	Device 2	Jammer	38500	40380	Diyarbakir
4	Device 3	Radar	38500	40390	Diyarbakir

Table 3.4 Failure rates according to bit results

Battery St.	Battery Lvl.	Heat St.	Connection St.	System St.	Failure
1	[2,5]	0	0	0	0
1	[2,5]	1	0	0	50% 0; 50% 1
1	[2,5]	1	0	1	30% 0; 70% 1
1	[2,5]	0	1	1	50% 0; 50% 1
1	[2,5]	1	1	1	10% 0; 90% 1
0	[1,2]	0	0	0	10% 0; 90% 1
0	[1,2]	1	0	0	50% 0; 50% 1
0	[1,2]	1	0	1	20% 0; 80% 1
0	[1,2]	0	1	1	30% 0; 70% 1
0	[1,2]	1	1	1	10% 0; 90% 1

Table 3.5 Experiment results

Dataset	Model	Test accuracy (%)	Test accuracy after PCA (%)
Hand-crafted single device	SVM	96	96
Synthetic single device	SVM	76	78
Synthetic multiple device	MLPNN	80	86

3.5.2 Experiment Results

In this section, we ran the algorithms that we adapted to the datasets. Results are provided in Table 3.5.

In the tests performed on the HSD dataset, users were asked to obtain intuitive results regarding error occurrences, and we compared our test set with these results. Utilizing the PCA method could not boost the test results in that case. It only achieved the same result when choosing the *n_components* parameter (number of features) as 6, 7, or 8. Although the performance rate is high, the fact that it was done with a single device may have made it easier for the models to memorize the behaviors in the dataset.

In the test performed on the SSD dataset, the model was able to produce results quickly because the data set did not contain unexpected situations. Utilizing the PCA method upgraded the test results by 2%. Choosing the *n_components* parameter as 5, achieved a 78% accuracy while the original test result was 76%.

Finally, we performed our tests with the MLPNN algorithm on the SMD dataset. This data set is a more complex data set because it contains more than one device. Utilizing the PCA method upgraded the test results by 6%. Choosing the *n_components* parameter as 9, achieved a 86% accuracy while the original test result was 80%.

3.6 Conclusion and Future Work

In this study, we aimed to predict whether a jammer or radar device will break down or not for PHM usage. We created datasets for a field that has not been tried before and applied SVM and MLPNN algorithms. According to the results, prediction success rates were successful. In the future, this work can be taken further with unexpected error situations and deeper features in datasets taken from real devices. Further, some sensor data from the field such as vibration level, and dust level can be gathered. Extra features can be created through time series characteristics of our data such as “break down count”. Breakdown count is the number of breakdowns of the same device in a time interval which can be driven from the record dates. One of the possible

limitations that may arise here is that most of the studies in this field are located in the military domain and therefore finding public dataset or device information is more difficult than standard studies.

References

1. Akhloufi, M.A., Arola, S., Bonnet, A.: Drones chasing drones: reinforcement learning and deep search area proposal. Drones **3**(3), 58 (2019)
2. Ayad, S., Terrissa, L.S., Zerhouni, N.: An iot approach for a smart maintenance. In: 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET), pp. 210–214. IEEE (2018)
3. Blackblaze: The backblzae hard drive data and stats. Technical report (2018). https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data
4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
5. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989)
6. Davari, N., Veloso, B., Costa, G.d.A., Pereira, P.M., Ribeiro, R.P., Gama, J.: A survey on data-driven predictive maintenance for the railway industry. Sensors **21**(17), 5739 (2021)
7. Guvenc, I., Koohifar, F., Singh, S., Sichitiu, M.L., Matolak, D.: Detection, tracking, and interdiction for amateur drones. IEEE Commun. Mag. **56**(4), 75–81 (2018)
8. Kazma, B., Semiz, F.: Radar/jammer device failure prediction datasets. Technical report (2024). https://kaggle.com/datasets/buketkazma/radarjammer-device-failure-prediction-datasets
9. Keipour, A., Mousaei, M., Scherer, S.: Alfa: a dataset for uav fault and anomaly detection. Int. J. Robot. Res. **40**(2–3), 515–520 (2021)
10. Khalil, K., Eldash, O., Kumar, A., Bayoumi, M.: Machine learning-based approach for hardware faults prediction. IEEE Trans. Circuits Syst. I Regular Papers **67**(11), 3880–3892 (2020)
11. Lei, Y., Li, N., Gontarz, S., Lin, J., Radkowski, S., Dybala, J.: A model-based method for remaining useful life prediction of machinery. IEEE Trans. Reliab. **65**(3), 1314–1326 (2016)
12. Lei, Y., Lin, J., He, Z., Zuo, M.J.: A review on empirical mode decomposition in fault diagnosis of rotating machinery. Mech. Syst. Signal Process. **35**(1–2), 108–126 (2013)
13. Multerer, T., Ganis, A., Prechtel, U., Miralles, E., Meusling, A., Mietzner, J., Vossiek, M., Loghi, M., Ziegler, V.: Low-cost jamming system against small drones using a 3d mimo radar based tracking. In: 2017 European Radar Conference (EURAD), pp. 299–302. IEEE (2017)
14. Noh, J., Kwon, Y., Son, Y., Shin, H., Kim, D., Choi, J., Kim, Y.: Tractor beam: safe-hijacking of consumer drones with adaptive gps spoofing. ACM Trans. Privacy and Secur. (TOPS) **22**(2), 1–26 (2019)
15. Park, S., Kim, H.T., Lee, S., Joo, H., Kim, H.: Survey on anti-drone systems: components, designs, and challenges. IEEE Access **9**, 42635–42659 (2021)
16. Shi, X., Yang, C., Xie, W., Liang, C., Shi, Z., Chen, J.: Anti-drone system with multiple surveillance technologies: architecture, implementation, and challenges. IEEE Commun. Mag. **56**(4), 68–74 (2018)
17. Takashima, T., Yamaguchi, J., Otani, K., Kato, K., Ishida, M.: Experimental studies of failure detection methods in pv module strings. In: 2006 IEEE 4th World Conference on Photovoltaic Energy Conference. vol. 2, pp. 2227–2230. IEEE (2006)
18. Veloso, B., Gama, J., Ribeiro, R.P., Pereira, P.M.: A benchmark dataset for predictive maintenance (2022). arXiv:2207.05466
19. Wei, S., Yuying, L., Minjie, D., Sai, Z.: Radar equipment fault diagnosis method based on support vector domain description. In: 2011 International Conference on Electronics, Communications and Control (ICECC), pp. 2193–2195. IEEE (2011)

20. Wellig, P., Speirs, P., Schuepbach, C., Oechslin, R., Renker, M., Boeniger, U., Pratisto, H.: Radar systems and challenges for c-uav. In: 2018 19th International Radar Symposium (IRS), pp. 1–8. IEEE (2018)
21. Wu, D., Jennings, C., Terpenny, J., Kumara, S., Gao, R.: Cloud-based parallel machine learning for prognostics and health management: a tool wear prediction case study. J. Manuf. Sci. Eng. **140**(4) (2017)
22. Xie, W., Wang, L., Bai, B., Peng, B., Feng, Z.: An improved algorithm based on particle filter for 3d uav target tracking. In: ICC 2019-2019 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2019)
23. Yaacoub, J.P., Noura, H., Salman, O., Chehab, A.: Security analysis of drones systems: Attacks, limitations, and recommendations. Internet of Things **11** (2020)
24. Yasmine, G., Maha, G., Hicham, M.: Survey on current anti-drone systems: process, technologies, and algorithms. Int. J. Syst. Syst. Eng. **12**(3), 235–270 (2022)
25. Zhang, Y., Ma, Y.: Design of the bit base on the structure of the radar system. In: Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing). pp. 1–5. IEEE (2012)
26. Zonta, T., Da Costa, C.A., da Rosa Righi, R., de Lima, M.J., da Trindade, E.S., Li, G.P.: Predictive maintenance in the industry 4.0: a systematic literature review. Comput. Ind. Eng. **150**, 106889 (2020)

Chapter 4
Intelligent Augmented Reality System for Optimal Object Placement

Bianca-Ştefana Popa and Cosmin-Iulian Irimia

Abstract Home furnishing requires meticulous attention to detail and knowledge in design and functionality. Current platforms often lack intuitive visualization tools for furniture placement within a real space. Given the limitations observed in existing apps, this paper introduces an innovative Android application that utilizes augmented reality (AR). The app provides an intuitive environment, letting the user scan their room, make a furniture selection, and visualize it into an optimally designed arrangement, in real life, by using the smartphone's camera. This paper outlines the app's structure and its utility, implementing interior design constraints defined through mathematical functions, heuristics, and optimization techniques like *Backtracking* and *Simulated Annealing* algorithms, integrated with the AR system and 3D models to offer the most suitable furniture layout.

4.1 Introduction

Apartment explorer, an AR-based application, addresses the challenge of furniture arrangement by minimizing manual input and guesswork. Unlike existing tools, it provides optimal arrangement suggestions based on design principles and clear AR visualization. It stands out from apps like IKEA place, which doesn't allow simultaneous visualization of multiple items, and Amazon's offerings, where manual positioning is often required. Focused on home decoration, Apartment explorer auto-generates furniture arrangements [1] considering room dimensions, windows, and doors, enhancing user experience beyond typical AR-enhanced shopping apps. Users can scan their room, identify key features, and select from 3D furniture models, with the app visualizing the best arrangement. The following sections explore the algorithms and methodologies enabling this user-friendly and efficient solution.

B.-Ş. Popa · C.-I. Irimia (✉)
Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iasi, Romania
e-mail: cosmin.irimia@uaic.ro

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_4

4.2 Architecture

4.2.1 Technologies

The application was developed using Unity, chosen for its prominence in game and AR application development. Unity primarily utilizes C# for scripting, backed by comprehensive documentation and online resources. AR integration was achieved through Unity's AR Foundation framework. This framework supports *ARCore* for Android AR applications, offering features like six degrees of freedom for movement tracking, real-world surface detection, and AR session management. A key component, *ARRaycastManager*, generates virtual rays to detect surfaces and objects in AR, facilitating user interaction within the virtual environment.

The application features a selection of self-made 3D furniture models for virtual arrangement, created using Blender for design flexibility. This diverse yet concise collection covers essential items for bedrooms, living rooms, and kitchens, enabling users to customize their virtual spaces. The models were designed with reference to real-world dimensions for accuracy while maintaining simplicity to ensure smooth rendering in the app.

4.2.2 User Flow

The application consists of four scenes: *Main Menu*, *Room Scan*, *Choose Items*, and *Visualize Room*. In the *Main Menu*, users can create a new room, proceeding to the *Room Scan* scene. Using AR Foundation's plane detection, users move their phones to accurately capture the floor plan. Real-life objects may create additional unwanted detection, leading to irregular or overlapping planes. To solve this, two buttons have been implemented to facilitate the scanning process:

- *Clean Planes* disables smaller planes, retaining the largest one to address overlapping planes, as can be seen in Fig. 4.1a and b.
- *Reset Planes* clears all detected planes and placed objects (corners, doors, windows), useful when the detection doesn't match the floor plan.

An empty room is recommended for optimal object viewing, although not mandatory, as the 3D model positioning doesn't consider real-life objects obstructing the way. If the room is empty, the two helper buttons may not be necessary.

After detection, room corners are placed **consecutively**. This step ensures precise wall definition since AR-detected planes may not perfectly delineate the edges. A room must have four corners; extra corners will be ignored.

Users then choose to add doors or windows. Objects (corner, door, window) can be toggled on the screen's left, with the corner being default. In Fig. 4.1c–e there are displayed placement examples for each object. After choosing an object, the user taps on the screen to place it. Misplacements can be undone using the *Undo*

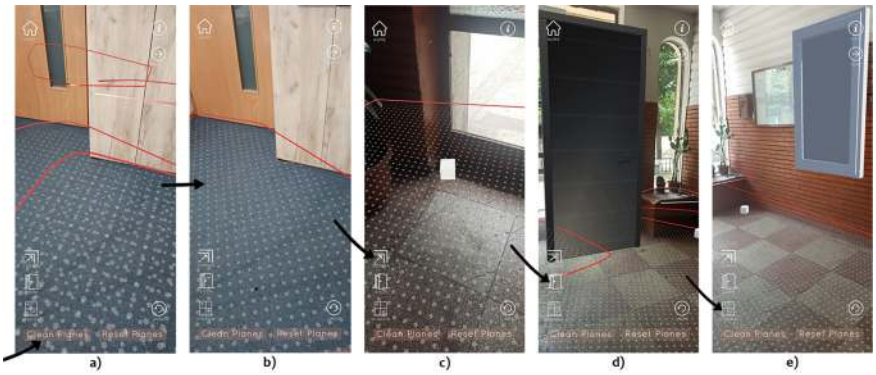


Fig. 4.1 **a**, **b** Clean planes, **c**–**e** Add corners, doors, windows

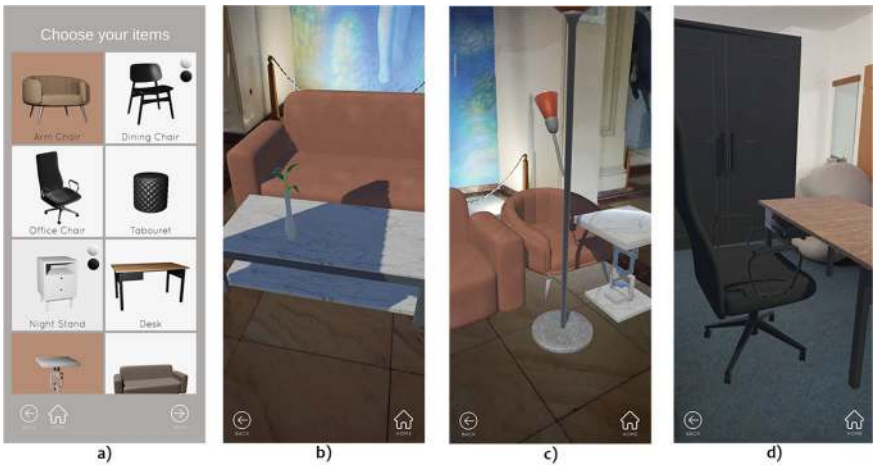


Fig. 4.2 **a** Choose items scene, **b**–**d** Visualize room scene

button. To avoid layout mismatches, door and window placements are restricted to the room's edges. Their position and rotation are adjusted to be parallel to the nearest wall, easing the placement process. The *Info* button provides a brief description of the scanning process.

Upon scanning completion, users can proceed to the *Choose Items* scene via the *Next* button. The furniture list is displayed as a two-column grid, each box showing the object's image and name, as can be seen in Fig. 4.2a. Some objects have color options.

In the final scene, selected objects are automatically positioned and viewable through the phone's camera in reality. Some examples of layouts are displayed in Fig. 4.2c–e. Users can navigate back to previous scenes to adjust parameters or selections using the *Back* button, enhancing the application's flexibility and user control.

If no object positioning is found due to space constraints, a warning message will be generated. The flexible and adaptive navigation ensures user comfort and control, contributing to an intuitive, enjoyable experience and maximizing the application's overall efficiency.

4.3 Defining the Problem

The application addresses the NP-hard problem of furniture placement, due to numerous variables, like potential positions and orientations of furniture and the many constraints that must be met.

The user-placed room corners, doors, and windows are saved as 2D positions, combined with selected furniture to create a "Layout" class instance. Furniture items are characterized within the "Item" class by their name, corners coordinates, orientation, as well as length and width. Their orientation is aligned relative to the room's initial wall, either widthwise or lengthwise.

The initial step in furniture arrangement involves identifying potential object positions within the room's polygon, which may not align with the x and y axes. This is done by forming a rectangle around the room's corners and iterating through coordinates using the *Ray Casting* Algorithm [2] to confirm polygon inclusion.

The application then calculates the object's corner positions, rotating them based on the angle relative to the room's first wall. Ensuring the object fits within the room involves checking that its first three corners remain inside the polygon. Position validation also includes confirming no overlap with existing furniture, doors, or windows, using an imaginary rectangle to maintain clearance. Overlap checks employ the *Separating Axis* Theorem [3], particularly for cases where object edges aren't parallel to the coordinate axes, verifying non-intersection by finding a separating line between two polygons.

4.3.1 Constraint Satisfaction Problem

Constraint Programming (CP) [4] is a powerful paradigm for solving complex combinatorial problems by defining constraints on variables and searching for a solution that satisfies all these constraints. Constraint satisfaction problems involve a set of objects assigned specific states and a set of constraints restricting their states. In the context of this application, the state of furniture items is determined by their location and orientation within the room space. Constraints are rules that determine where furniture pieces can be positioned:

1. **Door Positioning Constraint**: Furniture must not be positioned in front of doors to ensure uninterrupted access to and from the room.

2. **Window Positioning Constraint**: Certain items, such as those that block light, should not be positioned in front of windows.
3. **Wall Positioning Constraint**: Some pieces of furniture, like cabinets or desks, are typically positioned along walls. This approach is not only aesthetically pleasing but also optimizes available space.
4. **Relational Positioning Constraint**: Some objects are logically paired together, like a desk and an office chair, a bed and bedside tables, or a sofa and a coffee table. This relationship is described as parent-child.
5. **Space Constraint**: Each piece of furniture requires a specific amount of space around it to be utilized efficiently.
6. **Circulation Constraint**: This ensures there is comfortable movement within the room by maintaining suitable distances between various objects.

The first two constraints are considered essential and cannot be violated. Therefore, they are evaluated before a position is added to the list. The possible positions of an object are filtered in advance to ensure that they satisfy these two constraints.

4.4 Implementation and Algorithms

4.4.1 Details About AR

Object placement in the augmented reality scene, during the room scanning phase, is executed by a script that continuously monitors the user's screen touches throughout the application's execution. Upon detecting a touch, a ray is cast into the scene using the *ARRaycastManager* component. If this ray intersects with a flat surface, a 3D object is instantiated at the collision point, based on the user's selected option (corner, door, or window).

Moreover, during the room visualization which is managed by a separate script– the selected objects are automatically placed using data provided by the *Layout* class. Proper object orientation is another crucial aspect meticulously managed. The object's direction (forward/backward, left/right) is pre-calculated. The wall closest to the object is identified and the object's center is projected onto the opposite wall. The object is then rotated to face that point, utilizing Unity's *object.transform.LookAt()* function. This ensures that objects near a wall will be positioned with their backs against it, reflecting a realistic furniture arrangement in a room.

Another challenge is performance management, as augmented reality typically involves processing large amounts of data in real time. The 3D objects created are modeled to be not overly complex while retaining a semblance of realism.

4.4.2 Constraints Data

The constraints related to the spatial positioning of objects, discussed in previous sections, are defined using data structures, allowing for their subsequent manipulation. They are vital to ensure that objects are placed logically and usefully. One such data structure is the list *NearWallItems*. It enumerates objects to which the *Wall Positioning Constraint* applies.

ParentChildRelationships defines the *Relational Positioning Constraint*. It is structured as a dictionary where the keys contain the names of the parent and child objects, and the values represent the sides of the parent object where the child objects can be placed. For a sofa, multiple relationships are defined since the positioning of child objects relative to the parent can vary (a coffee table is specified to be in front of a sofa, while a side table or a floor lamp can be placed to either side of the sofa).

The dictionary *Clearance* refers to the *Space Constraint*. It enumerates, for specific objects, the sides that require space and the amount needed. The keys are tuples containing the object's name and a string with the sides of the object where space is applied. The values express the size of the required space, either relative to the object's own dimensions or those of its children. For example, a cabinet needs half of its length in front of it and a sofa needs space in front, left, and right of the width or length of his child.

4.4.3 Constraint Satisfaction Heuristic

The problem is solved using two types of algorithms. First, a *Backtracking* algorithm is used for generating an initial solution, which will then be optimized using the second algorithm, *Simulated Annealing*. Both utilize the *Constraint Satisfaction Heuristic*, often called a *prioritization* heuristic, used in search and optimization problems. It prioritizes the search process based on an estimated *benefit* of a solution, referred to as *score* or *reward*.

In this case, the heuristic is multi-criteria, considering multiple constraints when determining the quality of a position. Each constraint is associated with a reward, and the total *reward* is used to rank the solutions. The reward is calculated considering previously defined furniture positioning constraints. Each fulfilled constraint increases the *reward* by a certain value, depending on its importance. The heuristic function utilizes the data structures explained earlier, along with various mathematical functions:

- The **Wall Positioning Constrain**t offers a reward if the object is closer than 0.2 to the nearest wall.
- The next checked constraint is **Relational Positioning**. Initially, it checks if the current object is a child and if its parent was selected by the user. The reward is given in two steps, reflecting the importance of proper object placement:

  - Proximity to Parent Object: A reward is calculated proportional to the inverse distance between the current (child) object and the parent object, encouraging closer placement to the parent - the smaller the distance, the larger the reward.
  - Correct Positioning Relative to Parent: In the second step, if the child object is close enough to the parent (distance smaller than 0.3) and is on the correct side, an additional reward is added to the total.

  This reward method encourages not only the proximity of child objects to the parent but also their correct relative positioning. The reward is given only to the first found parent, being the most relevant.
- For the **Space Constraint**, the *Clearance* dictionary is referenced to check if the current object does not violate the needed space for the other selected objects. The object's extended position is calculated by adding the necessary space around it, as specified in the dictionary. Then, it's checked if this extended space overlaps with the current object. If not, a reward is added to the total. Thus, the current object receives a number of points for each object it doesn't overlap with. For some objects of greater importance, the reward increases. This approach promotes placement that maintains required clearance around objects for functional use, like ensuring unoccupied space for a cabinet's doors to open.
- The last constraint, **Circulation**, applies to larger objects and offers a reward proportional to the distance between them, encouraging more open space in the room.

4.4.4 Backtracking Algorithm

Backtracking (BKT) is a general algorithm used for finding all (or some) solutions to problems by incrementally constructing candidate solutions and abandoning a candidate ("backtracking") as soon as it establishes that it cannot possibly be completed into a full solution.

The BKT algorithm prioritizes larger objects by size and sequentially attempts to place each within a layout. It halts and returns true when all objects are placed. If placement fails, the algorithm backtracks and tries different positions. It ranks potential positions using the *Constraint Satisfaction Heuristic*, placing objects and recursively calling itself for the next object. The process repeats until it either finds a successful arrangement for all objects or exhausts all possibilities, returning false.

This approach can also be seen as a form of *Greedy* heuristic, where at each decision point (placing a furniture piece), the decision that seems best according to the heuristic is chosen. This doesn't always lead to the best overall solution as it doesn't consider the implications of its choices on future decisions. However, this approach can often provide good solutions.

The decision to use the BKT algorithm is tied to the unique specifications of the problem. For a constraint satisfaction problem, a BKT algorithm, adapted with a

Greedy behavior, offers a suitable result as it always tries to make the optimal choice at every step based on available information. While BKT can be time consuming for problems with a large solution space, its greedy variant is more efficient. It aims to find a solution quickly without exploring all possibilities, crucial for the problem at hand where a rapid and accurate initial layout solution is desired.

Therefore, the major advantage of using the BKT algorithm in this context isn't necessarily to provide the final solution but rather to generate an initial feasible arrangement for the *Simulated Annealing* algorithm. Starting from a layout that already satisfies the basic constraints can enhance the efficiency of the optimization process.

4.4.5 Simulated Annealing Algorithm

Considering the nature of the problem and the involved constraints, it is reasonable to explore alternatives to the standard *Backtracking* algorithm. *Constraint Satisfaction Problems* (CSPs) can often be more efficiently solved using other algorithms, such as local search or optimization algorithms.

One option is employing a stochastic local search algorithm, like *Simulated Annealing*(SA), which combines both greedy and random moves. This kind of algorithm can handle complex optimization problems and escape local minima by occasionally accepting worse solutions.

For implementing the SA algorithm, the existing *Constraint Satisfaction Heuristic* function is used as a fitness function to evaluate the quality of solutions. The algorithm will seek a solution that maximizes the fitness function while satisfying the constraints.

The SA function initializes a temperature and gradually cools it. At each iteration, it generates a new neighboring layout(calculated by the *Generate Neighboring Layout* function) and evaluates its state, using the *Evaluate Fitness* function. If the new layout is better, it is accepted. If not, it might still be accepted with a probability dependent on both temperature and fitness difference, using the *Accept Worse Solution* function.

The *Evaluate Fitness* function calculates the fitness of a layout by summing up the maximum rewards of all elements in the layout.

The *Accept Worse Solution* function calculates a probability using the exponential function, which will be higher if the difference between new and current fitness is large or if the temperature is high. A random number between 0 and 1 is generated. If this number is smaller than the calculated probability, the function returns true, accepting the worse solution; otherwise, it rejects the worse solution. This function allows the SA algorithm to escape from local minima by temporarily accepting worse solutions, hoping it leads to finding better solutions in the future. Temperature controls how often worse solutions are accepted. As the algorithm progresses, the temperature decreases, reducing the likelihood of accepting worse solutions.

The *Generate Neighboring Layout* function creates a new layout based on the current one by changing the position of a randomly selected item. The process varies depending on the importance of the selected item, as defined in the *Important Items* list. If the selected item isn't considered important, the function calculates all possible positions for that item within the current layout, selecting one randomly. If the selected item is important, the function temporarily removes all less important items from the layout, calculates all possible positions for the important item, selects one randomly, updates the item, and finally returns the less important items to the layout, also in randomly selected positions. This function provides a way to explore new possibilities for arranging items in the layout, allowing them to occupy positions that might otherwise be inaccessible due to the presence of other items.

SA offers several advantages over other optimization algorithms for this problem. One of its main benefits is its ability to escape local optima by occasionally accepting worse solutions. This feature allows the algorithm to explore a broader range of solutions, increasing its chances of finding the global optimum.

The algorithm is suitable for a wide range of optimization problems, including those with discrete search spaces like the furniture arrangement problem. It can also be easily adapted to work with various heuristics and constraints. Thus, simulated annealing is an appropriate choice for optimizing the furniture arrangement problem, which may involve complex constraints and a large search space with multiple local optima.

4.4.6 Examples

Figure 4.3a presents the arrangement of a set of living room furniture obtained by applying the SA algorithm, based on the previously described heuristic. The necessary space for door opening and other furniture-related constraints are met, ensuring easy access and comfortable circulation within the room. We can observe that, although the side table and the floor lamp have two potential parents, the sofa and the armchair, there exists a hierarchical priority between the parents, with the armchair being considered the more significant parent.

In Fig. 4.3b and c, two positions of a set of objects are displayed, illustrating that the algorithm can yield different results for the same input data set, a characteristic resulting from its **nondeterministic behavior**.

4.5 Algorithm Evaluation and Comparative Analysis

The algorithm's effectiveness primarily relies on its *Fitness* function, influenced by temperature and cooling rate parameters. Optimizing these parameters ensures a balance between solution quality and execution time, where a higher cooling rate decreases iterations and execution time at the expense of potentially less thorough

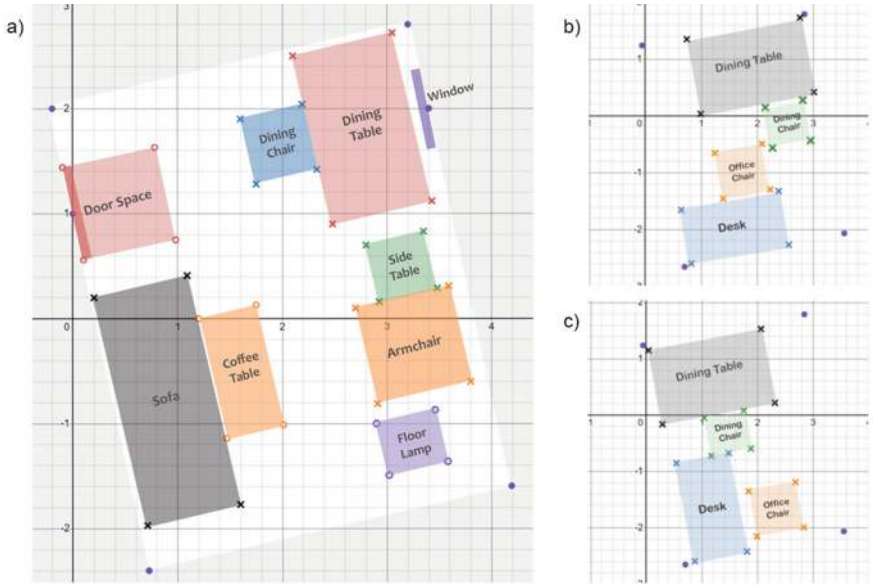


Fig. 4.3 Examples of layouts resulted from the algorithm

optimization. The solution's quality is further influenced by the reward system, which was established based on constraint importance, design principles, and empirical testing, enabling the algorithm to produce near-optimal arrangements with sufficient optimization iterations. Future enhancements could focus on refining the *Constraint Satisfaction Heuristic* used by the *Fitness* function to improve efficiency.

The algorithm's practicality is influenced by room size and object count; for instance, a 25 m^2 room with 5 objects takes 7–8 seconds, while 10 objects take up to 20 seconds, reflecting increased complexity. Despite variable execution times from its nondeterministic nature, the algorithm offers practical performance, given the problem's complexity and the high quality of its solutions.

When compared to a specific genetic algorithm approach for automated interior design [5], which optimizes layouts through design guidelines and mutations, the hybrid of *Simulated Annealing* and *Constraint Programming* excels by escaping the local optima, navigating complex spatial constraints more effectively and offering a superior solution quality and flexibility. The approach also establishes a dynamic balance between exploration and exploitation, an area where linear programming may fall short due to its deterministic nature. The discussed method thus offers a flexible and comprehensive solution to interior design's [6] complex demands.

4.6 Conclusion

Developing the application presented challenges in search algorithms, optimization, and augmented reality, with Simulated Annealing identified as the best fit. The application is now efficient and scalable, allowing easy addition of objects and intuitive, realistic 3D visualizations. Future improvements could include integrating furniture catalogs, saving functionalities, and algorithm optimizations. Overall, the project showcases artificial intelligence's role in enhancing interior design, offering a robust tool for furniture arrangement visualization.

References

1. Tang, J. K.T., Lau, W.-M., Chan, K.-K., To, K.-H.: 2014 International Conference on Virtual Systems and Multimedia (VSMM). AR interior designer: Automatic furniture arrangement using spatial and functional relationships, pp. 345–352
2. Ray Casting Algorithm. https://rosettacode.org/wiki/Ray-casting_algorithm. Accessed Mar 2023
3. The Separating Axis Theorem(SAT) (2010). https://dyn4j.org/2010/01/sat/. Accessed Feb 2023
4. Dirakkhunakon, S., Suansook, Y.: 2008 International Conference on Computer and Electrical Engineering. Stochastic Search Algorithm for Constraint Satisfaction Problem, pp. 682–686
5. Automated Interior Design Using a Genetic Algorithm. https://publik.tuwien.ac.at/files/publik_262718.pdf. Accessed Feb 2023
6. Vaidya, G.M., Loya, Y., Dudhe, P., Sawarkar, R., Chanekar, S.: 2022 Fifth International Conference on Computational Intelligence and Communication Technologies(CCICT). Visualization of Furniture Model Using Augmented Reality, pp. 488–493

Chapter 5
Mapping Research Publications Across the World: Looking for Opportunities in AI

Santiago Alonso, Abraham Gutiérrez, and Jesús Bobadilla

Abstract Understanding the presence of research opportunities in specific geographic areas is pivotal for research groups. This document, based on previous data mining within the field of artificial intelligence and subsequent data processing using machine learning techniques, presents an analysis by socioeconomic zones of scientific publication volumes, their corresponding fields, and potential relationships among them. It also delves into studying the correspondence between geographic areas and the quality of their production, assessed through publication impact factors and article citations.

5.1 Introduction

This article conducts an analysis of the work leading to the acquisition of data, subsequently enabling the presentation of diverse research areas based on geographical regions, thereby highlighting emerging trends. Consequently, this knowledge extraction allows for three potential courses of action: (a) identifying deficiencies in research within specific countries or geographical zones, prompting governments and organizations to invest in rectifying them; (b) pinpointing areas of opportunity to address these deficiencies through new research possibilities; and (c) identifying leading research areas within specific geographical zones that might prove interesting and/or profitable.

The work presented here is limited, by way of example, to the field of artificial intelligence, chosen for its inclusion in the Journal Citation Report (JCR) and for

S. Alonso (✉) · A. Gutiérrez · J. Bobadilla
ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, Calle Alan Turing S/N, 28031 Madrid, Spain
e-mail: santiago.alonso@upm.es

A. Gutiérrez
e-mail: abraham.gutierrez@upm.es

J. Bobadilla
e-mail: jesus.bobadilla@upm.es

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_5

encompassing disciplines that are present in research across all geographical regions. Our main contribution is to demonstrate that a study of opportunities worldwide can be conducted based on the analysis of scientific publications in any selected research area. To achieve this, artificial intelligence techniques are employed to determine the subareas to which each article belongs and a database is generated to establish the relationship between the different terms obtained and the geographical areas associated with these publications. This statistical study can be expanded in many ways, not only encompassing geographical zones but also considering different author teams or their affiliations.

A public link to the framework that allows the reconstruction of the database we have worked on and is valid to generate the dataset, is included at the following URL: http://salonso.etsisi.upm.es/idt/sd4aidb.sql.

In the line of trend and research gap analysis, various studies delve into the quality and quantity of publications across different scientific fields. Fontelo and Liu [1] reviews publication trends and identifies countries with the highest scientific output documented in PubMed from 1995 to 2015. Criticism regarding the quality and bias of existing reviews and meta-analyses is presented in [2], offering an analysis of their redundancies and conflicts of interest.

The critical analysis of publications considering the geographical factor spans various specialized fields. For instance, [3] examines the field of cardiology. A very recent study [4] scrutinizes publications in the field of artificial intelligence, focusing specifically on patents. A broader study [5] investigates patents across various technological domains.

Finally, the suitability of choosing Scopus as a bibliographic information source over Google Scholars is discussed. Martín-Martín et al. [6] conducts a current and detailed study comparing Scopus/WoS and Google Scholars, finding that "on average, citations from Google Scholars have much less scientific impact than those from Scopus/WoS."

The existence of public datasets for research purposes facilitates studies that rely on specific data samples. Consequently, numerous public datasets are available, such as those found in Google Trends, WHO data, etc., or well-known datasets like MovieLens [7] or FilmTrust [8]. However, identifying research areas being developed across different geographical regions requires a highly specialized dataset. The work conducted by [9, 10] provides a relevant dataset in the field of artificial intelligence.

In [10], the data mining phase gathers all necessary information from scientific articles within the field that have been written and submitted to the most significant journals worldwide. Scopus (Elsevier's) serves as the primary information source. This process yields a database containing the most relevant information, structured to include, among other details, the article title, its author, the number of citations, associated impact factor, author-indicated keywords, and document keywords.

The relevance of the articles is determined by four factors: (1) they are selected from Scopus, which holds worldwide recognition, (2) only articles falling exclusively within the top third of the first quartile (Q1) in the field of Computer Science—Artificial Intelligence are chosen, (3) each article is associated with its publication's impact factor, and (4) the number of citations for each article is available.

From the database, the authors construct a dataset structured as tuples: (paper, topic, and cardinality) through a process that determines the significance of each covered topic in the article, reflected in the "cardinality" aspect. The algorithm performing this is described in [10] and, as previously mentioned, it requires preprocessing that converts phrases into items. Utilizing this dataset, the authors can ascertain the article's affiliation with a sub-area within the field of artificial intelligence knowledge.

The matrix resulting from the tuple structure is comparable to any other dataset, with a sparsity of over 99%, making it a candidate for applying machine learning techniques to make predictions and recommendations to potential readers interested in a specific topic.

To verify the dataset's validity, a series of experiments are conducted where the quality of predictions and recommendations determines its suitability. These experiments are conducted using the Collaborative Filtering for Java (CF4J) environment [11]. This environment offers functionalities to load datasets designed for recommendation systems, select collaborative filtering methods and algorithms, execute processes, obtain predictions and recommendations, and measure result quality. Additionally, it allows the inclusion of new algorithms for comparison with the proposed references.

Once the environment is applied to the dataset obtained from the previous process, reference methods are employed to compare them with other datasets: (a) traditional statistical methods (Pearson correlation coefficient), (b) memory-based similarity measures: PIP [12], (c) Probabilistic Matrix Factorization method [13], and (d) methods based on neural networks [14, 15]. Mean absolute error is used to measure the prediction quality, while precision, recall, and normalized discounted cumulative gain measures are used to evaluate the recommendation quality.

Therefore, there is a valid dataset specifically suited to gather information enabling recommendations for researchers and authors regarding scientific documentation. There is also the possibility of conducting clustering techniques [16] on the dataset to perform classifications on the data.

Given this and recognizing the importance that descriptors (items or "topics") hold for managers and developers [17], processing of these descriptors is initiated. Due to the sparsity presented in the data matrix, matrix factorization [14] is utilized in collaborative filtering recommendation systems [18].

Using these techniques, in [9], the authors conduct a practical case analysis on scientific production based on research trends. This analysis enables the creation of a descriptor ranking that is subsequently grouped into areas through a clustering process. These areas gather associated topics within artificial intelligence.

Once the data quality, its significance, and appropriate structure are established, an analysis is conducted to provide recommendations to researchers aiming to capitalize on hidden opportunities or underexplored research areas. Furthermore, once adequate information is available, it becomes possible to identify burgeoning research areas in a specific region to potentially engage with them with a certain assurance of their "future."

For this reason, an analysis of the data focusing on geographical areas is presented, aiming to identify the research topics developed in each of them. This involves comparing the research carried out in different regions and forecasting potential future trends. The goal is also to compare the quantity and quality of production. There are studies that address measuring article quality based on a correlation between their quantity and the number of citations they receive [19].

5.2 Methodological Development of the Study

Initially, the relational database, as described in [10], undergoes modification to construct the dataset. The resulting structure comprises an extensive database consisting of 18 tables containing comprehensive information necessary for complete data processing.

A clustering process detailed in [9] leads to the identification of several research areas defined by the set of descriptors derived from distinct clusters. These groupings were subsequently reviewed by a recognized expert in the field. The expert, relying on his professional judgment, adjusted any inaccuracies stemming from the clustering process due to semantic incorporation.

As mentioned before, our work has focused on the field of artificial intelligence, even though it could be applied to any other area where a sufficient number of publications are available to ensure statistical correctness in the study. On the other hand, the foundation of the work lies in the accuracy of the available data, which is why Scopus has been chosen as a significant and reliable source. However, any bias in this information could significantly skew the results, even affecting prior processes like this clustering.

In Fig. 5.1 (production by countries), a notable disparity exists between China's production (6239) compared to other countries, where, at best (the United States: 2469). Regarding the European continent, a limited number of countries account for 70% of the production.

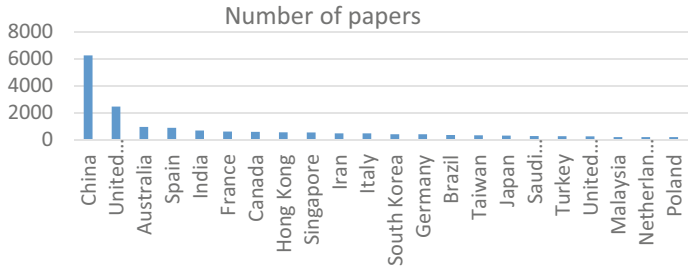


Fig. 5.1 Number of articles per country

Five geographical zones are chosen for this study: China, the United States, Oceania, India, and Europe. These geographical regions are selected for their significant scientific production and for constituting socioeconomic zones with individual characteristics. Initially, the number of articles published in each of these zones is calculated (Table 5.1). Once again, China's remarkably high participation is evident. This prompts an interest in analyzing the fields that receive the most attention in this country. Additionally, the combined contribution of European countries stands out as the second largest globally in these research fields, whereas other regions like Australia or India, with different socioeconomic situations, contribute considerably but incomparably less than the former. Nonetheless, their involvement as significant demographic zones remains notably prominent globally.

Firstly, it's worth noting that the production from the five selected geographical zones represents around 80% of global production. Secondly, observing Fig. 5.2 reveals how the number of publications in each area maintains the same proportion concerning global production. This is led by research areas labeled as "learning algorithms" and "information systems." The research fields at the bottom of the production scale are also consistent both globally and among the five studied zones: "recommender systems" and "social networks."

From Fig. 5.2, a preliminary conclusion can be drawn: if a researcher seeks a field within artificial intelligence to work in—internationally recognized, with a

Table 5.1 Number of articles per geographical area

Geo. area	Number art
China	6394
Europe	3831
USA	2469
Oceania	1026
India	696

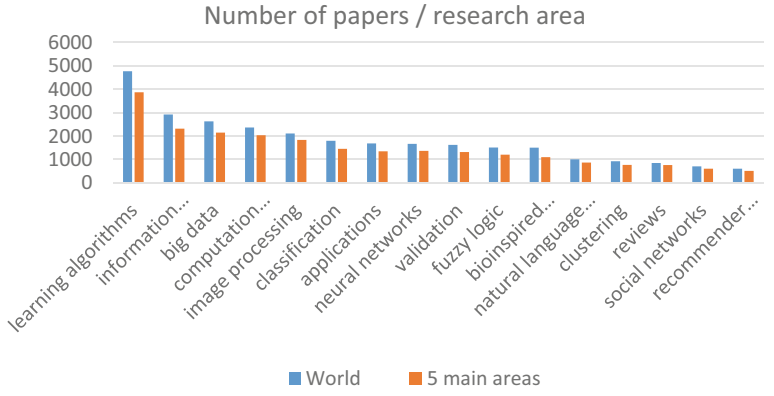


Fig. 5.2 Number of articles per research area

considerable number of international publications, yet not overly exploited—consideration should lean towards one of the fields situated in the last five positions of Fig. 5.2. Choosing one of these areas: "recommender systems," "social networks," or "reviews," among others, seems to ensure research opportunities. On the other hand, if the researcher aims to join a predominantly worked area, likely presenting funding opportunities and a greater number of groups working in it (thus more exploited), then selecting one of the first: "learning algorithms," "information systems," or "big data," among others, would be preferable.

The next natural step is to detect the different research areas showing higher production in each of the five geographical zones, allowing researchers to observe the differences and decide on areas of greater interest within their geographical zone. The most investigated topics in each of the zones are shown in Fig. 5.3.

Analyzing Fig. 5.3, the first thing that becomes evident is that the research area with the highest number of publications (thus arguably the most exploited area) is "learning algorithms" in all geographical zones. The current globalization, coupled with the ease of forming international research groups using current technical resources, may contribute to this being a globally exploited area. The difference between this area and the second one is notable, emphasizing its significant importance.

From the perspective of research areas, when looking at the proportion of the number of publications between geographical zones, the trend remains consistent in both areas with higher publication rates, such as "learning algorithms" (Fig. 5.4 left), and those with lower ones, like "social networks" and "recommender systems" (Fig. 5.4 right).

Evidently, the agreement in thematic areas where publications are predominantly focused across all geographical zones indicates a worldwide globalization concerning these topics. Differences between zones seem to refer not so much to the theme itself but to the quantity of publications. This likely correlates with the availability of resources to fund research and work leading to these publications, as well as the ability to finance the publications themselves.

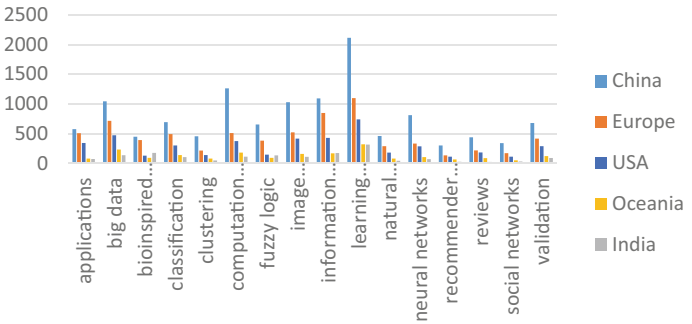


Fig. 5.3 Research areas per geographical areas

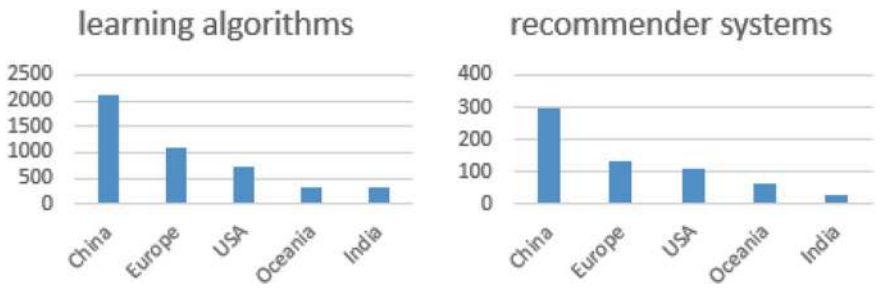


Fig. 5.4 Similar proportions between geographical areas in the research areas

5.3 Quality of Production

After analyzing the distribution of publications across geographical research areas, the next objective is to study the quality of this research. Naturally, a higher number of publications might imply a decrease in their quality. Given that this study focuses on the overall quality of research work across geographical zones, the quality index used should ideally refer to the three available factors: (1) the quantity of production, (2) the number of citations received from other articles, and (3) the impact factor of each publication.

The data regarding the quantity of production have already been presented and could potentially allow for establishing a ranking among the studied geographical zones.

Given that we have data regarding the number of citations for each article, we could establish a direct correlation between this number and its importance. After all, it reflects the importance that other researchers attribute to each article. By correlating the number of articles and the number of citations, we obtain the average citations per article in each of the zones (Table 5.3).

When studying Table 5.3 individually and considering the average number of citations as an index of quality, it can be observed that in this case, the geographical zone leading in this aspect is Oceania, with an average of 4.74 citations per published article, surpassing both China (4.39) and considerably exceeding the figures for the United States, Europe, and India. Although the numbers may not seem drastically different, this graph represents the fact that articles published in the Oceania zone

Table 5.3 Average number of citations per article

Area	Number of papers	Number of cites	Cites/paper
Oceania	1026	4924	4.74
China	6394	29,880	4.39
India	696	2636	3.79
USA	2469	8830	3.58
Europe	3831	15,837	3.48

Table 5.4 Average impact factor by geographical areas

Area	Avg. impact factor
Oceania	5.20
USA	4.97
China	4.78
Europe	4.60
India	4.16

receive 36% more citations than those published in Europe and 25% more than those published in the United States. These numbers suggest that works published in Oceania, despite being a smaller quantity compared to other zones, exhibit significant quality and serve as references in the field of artificial intelligence.

Finally, it's also possible to independently study the impact factor of publications where different articles appear. If an article has been published with multiple nationalities (authors from different countries) within the same geographical zone, the article's impact factor will be counted only once for that zone. Once calculated, the order described in Table 5.4 can be established.

As seen, once again, Oceania leads the series, surpassing the US and China. Naturally, this ranking, coupled with the previous one that referred to the number of citations received, speaks to the quality of the publications carried out by researchers in Oceania. Moreover, it seems reasonable to conclude that an article in a publication with a high impact level, i.e., in a higher quality journal, will be cited more frequently due to its inherent quality and the trust of other researchers. Therefore, a correlation between these two "indices" also seems reasonable.

Figure 5.5 presents the average impact factor in each research area, both globally and in Oceania. This comparison reveals that the average impact factor in all research areas in Oceania surpasses the average impact factor in the studied zones. Particularly noteworthy is that the impact factor of articles published in the "fuzzy logic" field in Oceania exceeds the overall average by almost one point. This establishes the articles published in this area in Oceania as a global reference within their field.

This graph also potentially guides researchers towards publications that could have a greater influence within the scientific community. Nonetheless, they are arranged in the graph (Fig. 5.5) based on their average global impact factor, providing an insight into the likely influence of a scientific article in a particular research area.

5.4 Conclusions

The aim of this article was to continue the authors' previous study concerning publications in the field of artificial intelligence, allowing for the construction of recommendation systems that analyze the similarity between scientific articles to provide

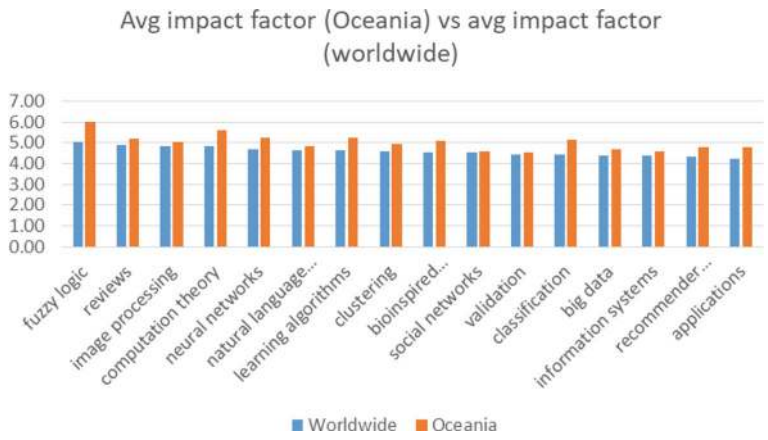


Fig. 5.5 Comparison of the average impact factor of Oceania with the world average

recommendations to researchers. The obtained data enables a geographical analysis to determine themes of particular relevance in different areas, revealing the differences between them.

These data allow for a quantitative analysis that, once conducted, has determined the most productive regions in terms of the number of publications, with China standing out prominently. Naturally, these results are associated with various factors such as larger population, greater economic power, or even just a higher research budget.

In the quantitative examination, the participation of each geographical zone studied is proportional across almost all different research areas. It seems that the aspect of "globalization," with its accompanying ease of access to resources and collaboration with research teams worldwide, homogenizes this aspect. It's noteworthy that the two most prominent research areas in all geographical zones are "learning algorithms" and "information systems."

Regarding qualitative analysis, the various available indicators—number of publications, citations received, and publication impact factor—show a slight change in the ranking of geographical zones. Oceania stands out as a region where scientific production in the field of artificial intelligence receives a noteworthy number of citations while being published in journals with the highest impact factor. This demonstrates a high level of quality attributed to both by editors and the global research community.

Referring to research areas, a ranking of topics has been obtained based on their "influence" in the scientific community according to their average impact factor.

Lastly, it's crucial to emphasize the importance of conducting these studies, in all their steps, for all fields of science where there's a need to identify trends and opportunities.

Acknowledgements This work was partially supported by the Ministerio de Ciencia e Innovación of Spain under the project PID2019-106493RB-I00 (DL-CEMG) and the Comunidad de Madrid under the Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of the Programa de Excelencia para el Profesorado Universitario.

References

1. Fontelo, P., Liu, F.: A review of recent publication trends from top publishing countries. London: BMC Syst. Rev. **7**(1), 147–147 (2018)
2. Ioannidis, J.P.: The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. **94**, 485–514 (2016). https://doi.org/10.1111/1468-0009.12210
3. Pagel, P.S., Hudetz, J.A.: A bibliometric analysis of geographic publication variations in the journal of cardiothoracic and vascular anesthesia from 1990 to 2011. Philadelphia: Elsevier Inc. J. Cardiothorac. Vasc. Anesth. **27**(2), 208–212 (2013). https://doi.org/10.1053/j.jvca.2012.08.022
4. Jiang, L., Chen, J., Bao, Y., Zou, F.: Exploring the patterns of international technology diffusion in AI from the perspective of patent citations. Scientometrics (2021). https://doi.org/10.1007/s11192-021-04134-3
5. Yang, W., Yu, X., Zhang, B., et al.: Mapping the landscape of international technology diffusion (1994–2017): network analysis of transnational patents. J. Technol. Transf. **46**, 138–171 (2021). https://doi.org/10.1007/s10961-019-09762-9
6. Martín-Martín, A., Orduna-Malea, E., Thelwall, M., Delgado, L.-Cózara. E.: Google scholar, web of science, and Scopus: a systematic comparison of citations in 252 subject categories. J. Informet.Informet. **12**(4), 1160–1177 (2018)
7. Harper F.M., Konstan J.A.: The MovieLens datasets: history and context. Proc. ACM Trans. Interact. Intell. Syst. **5**, 19:1–19:19 (2015)
8. Golbeck, J., Hendler, J.: FilmTrust: movie recommendations using trust in web-based social networks. In: Proceedings of 3rd IEEE Consumer Communications and Networking Conference (CCNC), vol. 1, pp. 282–286 (2006)
9. Bobadilla, J., Gutiérrez, A., Patricio, M.A., Bojorque, R.X.: Analysis of scientific production based on trending research topics. In: An Artificial Intelligence Case Study. Revista Española de Documentación Científica (2019)
10. Ortega, F., Bobadilla, J., Gutiérrez, A., Hurtado, R., Li, X.: Artificial intelligence scientific documentation dataset for recommender systems. IEEE Access **6**, 48543–48555 (2018). https://doi.org/10.1109/ACCESS.2018.2867731
11. Ortega, F., Zhu, B., Bobadilla, J., Hernando, A.: CF4J: collaborative filtering for java. Knowl. Based Syst. **152**, 94–99 (2018)
12. Ahn, H.J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Inf. Sci. **178**(1), 37–51 (2008)
13. Bokde, D., Girase, S., Mukhopadhyay, D.: Matrix factorization model in collaborative filtering algorithms: a survey. Procedia Comput. Sci. **49**, 136–146 (2015)
14. Xue, H.J., Dai, X.Y., Zhang, J., Huang, S.: Deep matrix factorization models for recommender systems. IJCAI 3203–3209. Melbourne, Australia (2017). https://doi.org/10.24963/ijcai.2017/447
15. Bobadilla, J., Ortega, F., Gutiérrez, A., Alonso, S.: Classification-based deep neural network architecture for collaborative filtering recommender systems. Int. J. Interact. Multimed. Artif. Intell. **6**(1) (2020)
16. Omran, M., Engelbrecht, A., Salman, A.: An overview of clustering methods. Intell. Data Anal. **11**, 583–605 (2007). https://doi.org/10.3233/IDA-2007-11602

17. Hindle, A., Bird, C., Zimmermann, T., Nagappan, N.: Do topics make sense to managers and developers? Empir. Softw. Eng.. Softw. Eng. **20**(2), 479–515 (2015). https://doi.org/10.1007/s10664-014-9312-1
18. Bobadilla, J., Ortega, F., Hernando, A., Gutierrez, A.: Recommender systems survey. Knowl. Based Syst. **46**, 109–132 (2013). https://doi.org/10.1016/j.knosys.2013.03.012
19. Hayati, Z.: Correlation between quality and quantity in scientific production: a case study of Iranian organizations from 1997 to 2006. Scientometrics **80**(3), 625–636 (2009). https://doi.org/10.1007/s11192-009-2094-3

Chapter 6
Diagnosis of Active Systems with Candidate Priority

Gianfranco Lamperti

Abstract Diagnosis of an *active system* (AS), an asynchronous and distributed discrete-event system, is typically *abduction-based*: given a temporal observation, the diagnoses, or *candidates*, are generated based on a complete model of the AS, where a candidate is a set of faults explaining the temporal observation. A critical problem, which is common to all approaches of model-based diagnosis, is a large number of candidates: this is a serious threat to diagnosticians, owing to the cognitive overload imposed by an overwhelming stream of information and, worse still, to the uncertainty raising from a large set of different diagnoses. This criticality is exacerbated by assuming that both the candidates and the relevant recovery actions, possibly performed by an artificial agent, are required in real time, like in a nuclear power plant or in a defense system. Since candidates with low cardinality are more probable than candidates with high cardinality, it seems appropriate to generate candidates in ascending order of cardinality, from most to least likely. This way, an agent is not required to wait for the complete generation of candidates to perform the recovery actions that are associated with the most probable diagnoses. A diagnosis technique for ASs with prioritization of candidates is presented. Evidence from experimental results shows that the diagnosis technique is not only sound and complete, inasmuch all and only correct candidates are generated, but also effective in providing the most likely candidates upfront.

6.1 Introduction

Diagnosis has always been a challenging task for Artificial Intelligence. Up to the mid-eighties, all approaches to diagnosis were *heuristics-based*: the knowledge of a human expert in a specific domain was embedded in a software system in form of rules mapping symptoms to possible diagnoses. No *deep* knowledge (i.e., knowledge on how the system works) was required, but only the experience of a diagnostician. Sub-

G. Lamperti (✉)
Department of Information Engineering, University of Brescia, Brescia, Italy
e-mail: gianfranco.lamperti@unibs.it

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_6

sequently, the diagnosis became *model-based*: a model of the normal behavior of the system is given in input to a diagnostic engine, which, based on a set of observations, generates the relevant diagnoses, called *candidates*, a candidate being a set of faulty components. Model-based diagnosis is grounded on the seminal works of Reiter [13], where a general theory of diagnosis is formalized in first-order logic, and de Kleer and Williams [6], where a *general diagnostic engine* is presented and experimented in the domain of troubleshooting digital circuits. In contrast with heuristics-based diagnosis, model-based diagnosis does not require any domain-dependent human expertise, but only a model of the *normal* (correct) behavior of the system: the discrepancy between the predicted (correct) behavior and the observed (faulty) behavior suffices to enumerate the candidates (diagnoses). The process for candidate generation requires two steps: (1) identification of the *conflict sets*, and (2) computation of the *hitting sets*. A conflict set is a set of components (e.g., devices in a digital circuit) such that, assuming that all of them behave correctly is logically inconsistent, in other words, at least one of them must be faulty: this is why this approach is also called *consistency-based*. Hitting sets are then generated from conflict sets in such a way that each hitting set intersects all conflict sets, that is, it includes at least one component from each conflict set. For instance, given the conflict sets $\{a, b\}$ and $\{a, c\}$, possible hitting sets are $\{a\}$ and $\{b, c\}$. Since the number of possible hitting sets may be very large, the diagnosis technique generates *minimal* hitting sets only: there is no hitting set that contains another hitting set. In our example, the hitting set $\{a, b\}$ is *not* minimal, as it includes $\{a\}$. Since the generation of both the conflict sets and the hitting sets are NP-hard problems, several techniques were proposed to make these two steps more efficient, possibly at the expense of completeness, including [1, 5, 16]. Assuming that the diagnoses are required under stringent time constraints, as is, for example, in a nuclear power plant or in a defense system, waiting for all the candidates may be less than desirable, as candidates with low cardinality (including few components) are generally more probable (hence, more valuable) than candidates with high cardinality (including many components). This is why it is paramount to the viability of a diagnostic system that candidates are generated in ascending order of cardinality, from most to least likely, so that a (possibly artificial) agent is enabled to perform effective recovery actions in real time, based on most realistic diagnoses [15, 17]. Consistency-based diagnosis is not the only technique adhering to the model-based paradigm: when time-varying (dynamical) systems come into play, *abduction-based* diagnosis may be a better choice, especially for discrete-event systems (DESs), which are the subject of a vast literature in the Control Theory [3]. Unlike the consistency-based approach, abduction-based diagnosis requires a *complete* model of the system, including both normal and faulty behavior. When the DES is distributed, each component is modeled as a finite automaton, where each state transition may be qualified either as normal or faulty. A trajectory of the DES, namely, a sequence of component transitions from the initial state of the DES to a final state, generates a sequence of observations associated with some (observable) transitions, namely a *temporal observation*, which is regarded as a symptom of the DES: the diagnostic engine may generate the candidates (sets of faults) by finding out the trajectories of the DES that conform with a given temporal observation, each

trajectory associated with a (not necessarily distinct) candidate. Abduction-based diagnosis of DESs is grounded on the seminal work of Sampath et al. [14], where the notion of *diagnosability* of a DES is coined and a *diagnoser* is defined in order to support the online diagnostic task efficiently. Other approaches to the diagnosis of DESs include [2, 4, 8–10, 12]. For the same reasons expressed above for consistency-based diagnosis, *minimal* diagnosis of DESs is proposed in [18], where minimal diagnosability is studied and a minimal diagnoser is proposed. What makes the construction of a diagnoser impractical for a real DES, however, is the need to first generate the space of the DES, whose number of states is exponential at least in the number of components. This is why, in this paper, we assume the unavailability of both the space and the diagnoser of the AS. We also assume that the diagnosis is required as soon as possible in order to be handled by an artificial agent designed to perform relevant recovery actions in real time. We propose a diagnostic engine where candidates are generated in ascending order of cardinality, from most to least likely, so that the most valuable diagnoses are generated upfront.

6.2 System Modeling

An active system is a network of *components* that are modeled as communicating automata. Each component is equipped with some input and/or output *pins*, where each output pin is connected with an input pin of another component by a *link*. A component changes its state by a *transition* that is triggered by an *event* either occurring outside of the system or coming from another component, after which the event is consumed, while other events may be generated on output pins. The newly generated events are placed on input pins of other components via links, thereby possibly triggering a cascade of additional transitions of different components. The behavior of an active system is assumed to be asynchronous: only one component transition at a time can occur.

Example 1 Depicted on the top-left of Fig. 6.1 is an active system $\mathcal{P}$ (protection) that is designed to control a cooling system by means of two components: a transducer z (incorporating a temperature sensor) and a valve v, with a link from z to v. In normal conditions, when the temperature becomes high, the transducer commands the valve to open in order to let the cooling fluid flow; vice versa, when the temperature returns to normal, the transducer commands the valve to close. Outlined on the bottom-left of the figure are the communicating automata of z (top) and v (bottom). Specifically, the model of z involves two states and four transitions, while the model of v involves two states and eight transitions. Generally speaking, each component transition from a state s to a state s' that is triggered by an input event e and generates a set of output events E is denoted by a triple $\langle s, (e, E), s' \rangle$. Component transitions in $\mathcal{P}$ are detailed in Table 6.1.

Starting from an initial state x_0, an active system reacts to a triggering event by a sequence of component transitions that moves the active system from x_0 to another

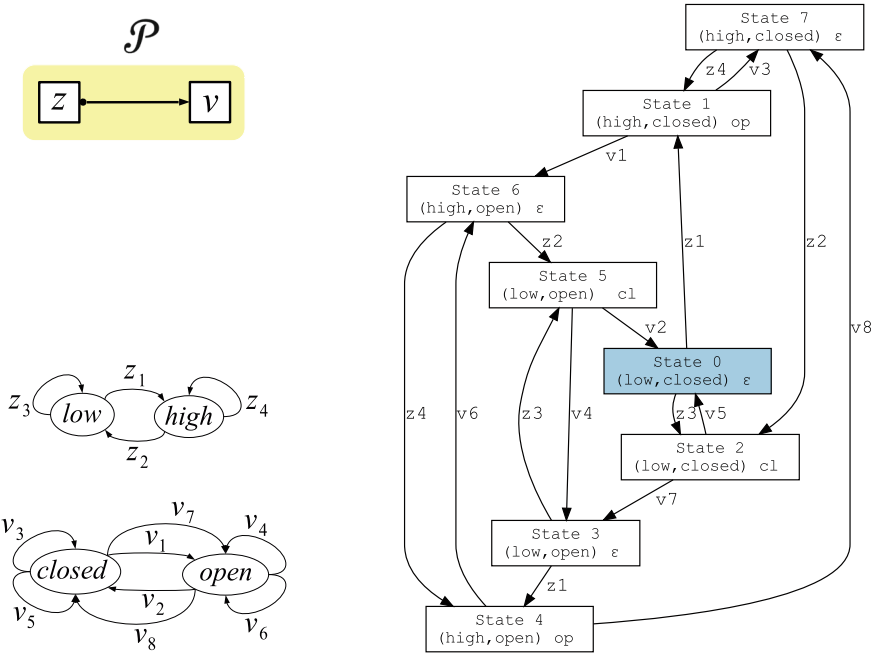


Fig. 6.1 Top-left: active system $\mathcal{P}$, involving transducer z and valve v; bottom-left: models of z (top) and v (bottom); right: $Space(\mathcal{P})$, where the (filled) initial state is 0

Table 6.1 Details of transitions of z and v (cf. component models in Fig. 6.1)

Component transition	Details
$z_1 = \langle low, (ko, \{op\}), high\rangle$	The transducer detects high temperature and generates the open event
$z_2 = \langle high, (ok, \{cl\}), low\rangle$	The transducer detects low temperature and generates the close event
$z_3 = \langle low, (ko, \{cl\}, low\rangle$	The transducer detects high temperature, yet generates the close event
$z_4 = \langle high, (ok, \{op\}), high\rangle$	The transducer detects low temperature, yet generates the open event
$v_1 = \langle closed, (op, \emptyset), open\rangle$	The valve reacts to the open event by opening
$v_2 = \langle open, (cl, \emptyset), closed\rangle$	The valve reacts to the close event by closing
$v_3 = \langle closed, (op, \emptyset), closed\rangle$	The valve does not react to the open event and remains closed
$v_4 = \langle open, (cl, \emptyset), open\rangle$	The valve does not react to the close event and remains open
$v_5 = \langle closed, (cl, \emptyset), closed\rangle$	The valve reacts to the close event by remaining closed
$v_6 = \langle open, (op, \emptyset), open\rangle$	The valve reacts to the open event by remaining open
$v_7 = \langle closed, (cl, \emptyset), open\rangle$	The valve reacts to the close event by opening
$v_8 = \langle open, (op, \emptyset), closed\rangle$	The valve reacts to the open event by closing

state x, called a *trajectory*. The (possibly infinite) set of trajectories of the active system is the language of a *finite* automaton, called the *space* of the active system.

Definition 1 The *space* of an active system $\mathcal{X}$ is a finite automaton,

$$Space(\mathcal{X}) = (\Sigma, X, \tau, x_0) \tag{6.1}$$

where the alphabet Σ is the set of component transitions, X is the set of states, where a state is a pair (C, L), with C being the array of states of components and L being the array of the (possibly empty) events within links[1], τ is the transition function mapping a state and a component transition into a new state, $\tau : X \times \Sigma \mapsto X$, and x_0 is the initial state.

Example 2 The space of $\mathcal{P}$ (cf. Example 1), namely $Space(\mathcal{P})$, is shown on the right side of Fig. 6.1, where boxes and arcs indicate states and transitions, respectively, with initial state 0. Each state, which is identified by a number in $0\,..\,7$, embodies the pair of the component states of z and v, as well as the (possibly empty) event in the link. Due to cycles in the space, there is an infinite number of trajectories of $\mathcal{P}$, such as $T = [z_3, v_5, z_1, v_3, z_4, v_3, z_2, v_5]$, which, incidentally, terminates in the initial state.

6.3 Observability and Abnormality

The specification of an active system as given is insufficient for diagnosis purposes: it needs to be augmented with information about the *observability* and the *abnormality* of the system, both of them being specified in a *mapping table*.

Definition 2 Let $\mathbf{T}$ be the set of component transitions in an active system $\mathcal{X}$, let $\mathbf{O}$ be a finite set of *observations* for $\mathcal{X}$, let $\mathbf{F}$ be a finite set of *faults* for $\mathcal{X}$, and let ε denote the *empty* symbol. The *mapping table* of $\mathcal{X}$ is a function:

$$Map(\mathcal{X}) : \mathbf{T} \mapsto (\mathbf{O} \cup \{\varepsilon\}) \times (\mathbf{F} \cup \{\varepsilon\}). \tag{6.2}$$

In practice, $Map(\mathcal{X})$ can be represented as a set of triples (t, o, f), with $t \in \mathbf{T}$, $o \in \mathbf{O} \cup \{\varepsilon\}$, and $f \in \mathbf{F} \cup \{\varepsilon\}$, where each triple defines the observability and abnormality of t, specifically: if $o \neq \varepsilon$, then t is *observable*, else t is *unobservable*; also, if $f \neq \varepsilon$, then t is *faulty*, else t is *normal*.

Example 3 The mapping table of active system $\mathcal{P}$ is listed on the left side of Table 6.2, where $\mathbf{O} = \{oz, ov\}$ and $\mathbf{F} = \{fz_3, fz_4, fv_3, fv_4, fv_7, fv_8\}$, which are detailed on the right side of the figure. Only one observation label is provided for both the transducer and the valve, namely oz and ov, respectively, each being associated with

[1] Formally, an empty link contains an empty event, denoted ε.

Table 6.2 Mapping table $Map(\mathcal{P})$ (left), and details of observations and faults (right)

t	o	f
z_1	oz	ϵ
z_2	oz	ϵ
z_3	ϵ	fz_3
z_4	ϵ	fz_4
v_1	ov	ϵ
v_2	ov	ϵ
v_3	ϵ	fv_3
v_4	ϵ	fv_4
v_5	ov	ϵ
v_6	ov	ϵ
v_7	ov	fv_7
v_8	ov	fv_8

o	*Observation details*
oz	The transducer performs a normal action
ov	The valve reacts (possibly abnormally) to an event

f	*Fault details*
fz_3	The transducer generates the *cl* event instead of *op*
fz_4	The transducer generates the *op* event instead of *cl*
fv_3	The valve remains closed upon the open command
fv_4	The valve remains open upon the close command
fv_7	The valve opens upon the close command
fv_8	The valve closes upon the open command

several (still not all) transitions. For instance, transition z_1 is observable and normal, z_3 is unobservable and faulty, whereas v_7 is both observable and faulty. Owing to the possible association of the same observation with *several* transitions, uncertainty still remains in determining the actual component transition based solely on the observation occurred.

According to a mapping table, each trajectory of an active system can be associated with both a *temporal observation* and a *diagnosis*.

Definition 3 Let T be a trajectory of an active system $\mathcal{X}$. The *temporal observation* of T is the sequence of the observations associated with the component transitions in T,

$$Obs(T) = \left[o \mid t \in T, (t, o, f) \in Map(\mathcal{X}), o \neq \varepsilon \right]. \tag{6.3}$$

A temporal observation $\mathcal{O}$ *conforms* with a trajectory T iff $\mathcal{O} = Obs(T)$.[2]

Example 4 According to the mapping table $Map(\mathcal{P})$ in Table 6.2 and considering the trajectory $T = [z_3, v_5, z_1, v_3, z_4, v_3, z_2, v_5]$, we have $Obs(T) = [ov, oz, oz, ov]$.

Definition 4 Let T be a trajectory of an active system $\mathcal{X}$. The *diagnosis* of T is the set of faults associated with the component transitions in T,

$$Dgn(T) = \{ f \mid t \in T, (t, o, f) \in Map(\mathcal{X}), f \neq \varepsilon \}. \tag{6.4}$$

A diagnosis δ *explains* a temporal observation $\mathcal{O}$ if $\delta = Dgn(T)$ and $\mathcal{O}$ conforms with T. The *cardinality* (number of faults) of δ is denoted $|\delta|$.

[2] The same temporal observation $\mathcal{O}$ may conform with several (possibly infinite) trajectories.

Example 5 Considering $Map(\mathcal{P})$ in Table 6.2 and $T = [z_3, v_5, z_1, v_3, z_4, v_3, z_2, v_5]$ in Example 2, we have $Dgn(T) = \{fz_3, fv_3, fz_4\}$.

Since a temporal observation $\mathcal{O}$ may conform with several trajectories of the active system, several (candidate) diagnoses may explain $\mathcal{O}$.

Definition 5 Let $\mathcal{O}$ be a temporal observation of an active system $\mathcal{X}$. The *candidate set* of $\mathcal{O}$ is the set of diagnoses of the trajectories of $\mathcal{X}$ conforming with $\mathcal{O}$,

$$\Delta(\mathcal{O}) = \{\delta \mid \delta = Dgn(T), T \in Space(\mathcal{X}), \mathcal{O} = Obs(T)\}. \tag{6.5}$$

Example 6 Let $\mathcal{O} = [ov, oz, oz, ov]$ be a temporal observation of active system $\mathcal{P}$. Based on $Space(\mathcal{P})$ in Fig. 6.1 and $Map(\mathcal{P})$ in Table 6.2, it is easy to find out that $\Delta(\mathcal{O})$ includes six diagnoses, or *candidates*, namely $\{fz_3, fv_3\}$, $\{fz_3, fz_4, fv_3\}$, $\{fz_3, fv_3, fv_7\}$, $\{fz_3, fz_4, fv_3, fv_7\}$, $\{fz_3, fv_3, fv_4, fv_7\}$, and $\{fz_3, fz_4, fv_3, fv_4, fv_7\}$.

6.4 Diagnosis Abduction

The candidate set of a temporal observation $\mathcal{O}$ of an active system $\mathcal{X}$ can be determined by generating a subspace of $Space(\mathcal{X})$ involving all and only the trajectories of $\mathcal{X}$ that conform with $\mathcal{O}$, called a *diagnosis abduction*.

Definition 6 Let $\mathcal{O} = [o_1, \ldots, o_n]$ be a temporal observation of an active system $\mathcal{X}$, with $Space(\mathcal{X}) = (\Sigma, X, \tau, x_0)$ and set of faults $\mathbf{F}$ involved in $Map(\mathcal{X})$. The *diagnosis abduction* of $\mathcal{X}$ based on $\mathcal{O}$ is a finite automaton,

$$Abd(\mathcal{X}, \mathcal{O}) = \left(\Sigma, A, \tau', a_0, A_f\right) \tag{6.6}$$

where A is the set of states, with each state being a triple (x, δ, i), where $x \in X$, $\delta \in 2^{\mathbf{F}}$, and $i \in [0\,..\,n]$ is an *index* of $\mathcal{O}$; $a_0 = (x_0, \emptyset, 0)$ is the initial state; $A_f \subseteq A$ is the set of final states, where the index i in each final state equals n (o_n being the last observation in $\mathcal{O}$); and $\tau' : A \times \Sigma \mapsto A$ is the transition function, where $\delta'((x, \delta, i), t) = (x', \delta', i')$ iff $\tau(x, t) = x'$, and, being (t, o, f) the triple in $Map(\mathcal{X})$: if $f = \varepsilon$ then $\delta' = \delta$ else $\delta' = \delta \cup \{f\}$, and, if $o = \varepsilon$ then $i' = i$, else, if $i < n$ and $o = o_{i+1}$, then $i' = i + 1$.

Example 7 Let $\mathcal{O} = [ov, oz, oz, ov]$ be a temporal observation of active system $\mathcal{P}$. Based on $Map(\mathcal{P})$ (Table 6.2), depicted in Fig. 6.2 is $Abd(\mathcal{P}, \mathcal{O})$, which includes 21 states (labeled $0, \ldots, 20$), where the initial state is 0, while the (twelve) final states are in bold. Each state is marked with a state in $Space(\mathcal{P})$ expressed as a pair of component states, a (possibly empty) event in the link, a vector of bits denoting a diagnosis, where each bit b_k, $k \in [1\,..\,6]$, indicates whether the k-th fault in the list $[fz_3, fz_4, fv_3, fv_4, fv_7, fv_8]$ is involved ($b_k = 1$) or not ($b_k = 0$) in the diagnosis, and an index $i \in [0\,..\,4]$ of $\mathcal{O}$. For instance, state 10 indicates that the states of z and v are

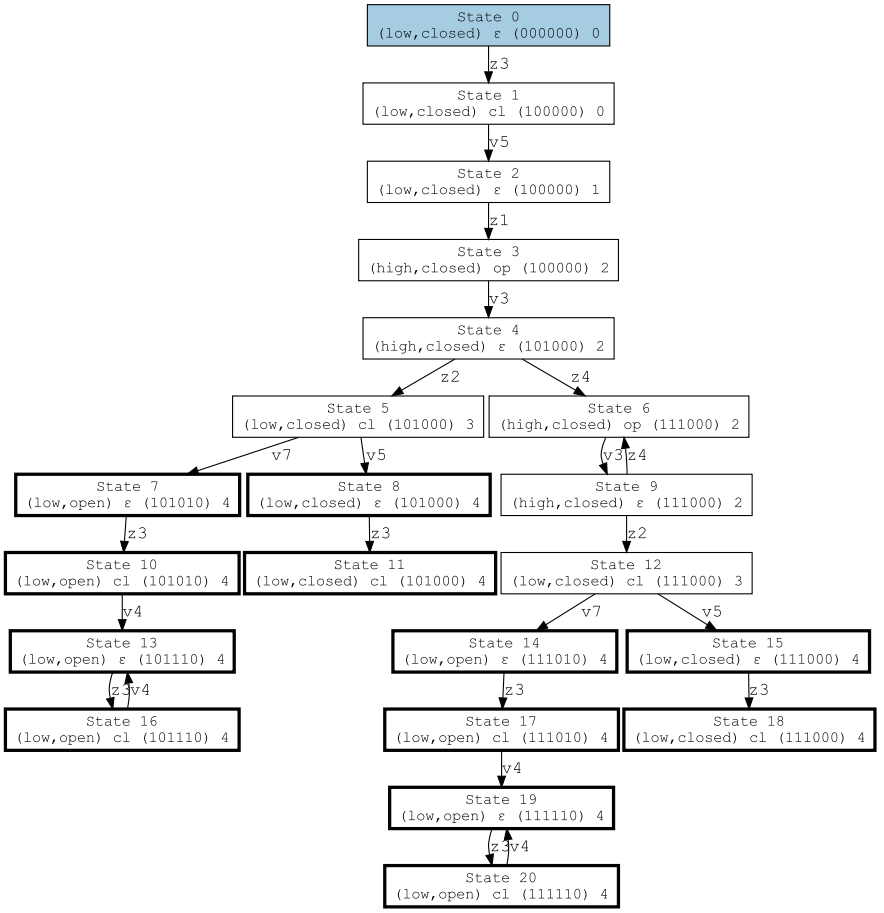


Fig. 6.2 $Abd(\mathcal{P}, \mathcal{O})$: the diagnosis abduction of $\mathcal{P}$ based on the temporal observation $\mathcal{O} = [ov, oz, oz, ov]$, with initial state 0 and final states being depicted in bold

low and *open*, respectively, the link contains event *cl*, the diagnosis is $\{fz_3, fv_3, fv_7\}$, and the index of $\mathcal{O}$ is $i = 4$ (hence, the trajectory ending in state 10 conforms with $\mathcal{O}$).

Proposition 1 *Let $\mathcal{O} = [o_1, \ldots, o_n]$ be a temporal observation of an active system $\mathcal{X}$, and let A_f be the set of final states in $Abd(\mathcal{X}, \mathcal{O})$. We have*

$$\Delta(\mathcal{O}) = \{\delta \mid (x, \delta, n) \in A_f\}\,. \tag{6.7}$$

Example 8 Based on $Abd(\mathcal{P}, \mathcal{O})$ in Fig. 6.2, the diagnoses associated with the final states are $\{fz_3, fv_3\}$ (states 8 and 11), $\{fz_3, fz_4, fv_3\}$ (states 15 and 18), $\{fz_3, fv_3, fv_7\}$ (states 7 and 10), $\{fz_3, fz_4, fv_3, fv_7\}$ (states 14 and 17), $\{fz_3, fv_3, fv_4, fv_7\}$ (states 13 and

16), and $\{fz_3, fz_4, fv_3, fv_4, fv_7\}$ (states 19 and 20), which, as claimed in Proposition 1, collectively constitute the candidate set $\Delta(\mathcal{O})$ determined in Example 6.

6.5 Prioritized Diagnosis Engine

Since not all candidates are equally probable, candidates are to be generated in ascending order based on their cardinality, on the grounds that the lower the cardinality, the more likely the candidate is the actual diagnosis (the diagnosis of the actual trajectory).

Definition 7 Let $\Delta(\mathcal{O}) = \{\delta_1, \ldots, \delta_m\}$ be a candidate set. An *ascending ordering* of $\Delta(\mathcal{O})$ is a sequence $\Delta^* = [\delta'_1, \ldots, \delta'_m]$ such that $\{\delta'_1, \ldots, \delta'_m\} = \{\delta_1, \ldots, \delta_m\}$ and $\forall k \in [1 \,..\, (m-1)] \left(|\delta'_k| \leq |\delta'_{k+1}|\right)$.

Example 9 Let $\Delta(\mathcal{O}) = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6\}$, where $\delta_1 = \{fz_3, fz_4, fv_3\}$, $\delta_2 = \{fz_3, fz_4, fv_3, fv_7\}$, $\delta_3 = \{fz_3, fz_4, fv_3, fv_4, fv_7\}$, $\delta_4 = \{fz_3, fv_3\}$, $\delta_5 = \{fz_3, fv_3, fv_7\}$, and $\delta_6 = \{fz_3, fv_3, fv_4, fv_7\}$. An ascending ordering of $\Delta(\mathcal{O})$ is $[\delta_4, \delta_1, \delta_5, \delta_2, \delta_6, \delta_3]$, where $|\delta_4| = 2$, $|\delta_1| = |\delta_5| = 3$, $|\delta_2| = |\delta_6| = 4$, and $|\delta_3| = 5$.

The diagnosis engine is expected to output Δ^* incrementally so that more probable candidates are generated *before* less probable ones. This way, a (software) agent designed to monitor the active system can possibly perform relevant recovery actions in real time. An algorithm, called PRIORITIZED DIAGNOSIS ENGINE, for the incremental generation of Δ^*, is listed below (lines 1–18).[3] The idea is to generate the abduction $Abd(\mathcal{X}, \mathcal{O})$ based on an ascending cardinality of the candidates, namely $\mathcal{C}$.[4] Only when the index of $\mathcal{O}$ is complete ($i = n$), is a candidate δ generated, where $|\delta| = \mathcal{C}$, provided it is not in Δ^* already (lines 5 and 6), thereby avoiding duplication of the same candidate. The loop in lines 7–15 generates the transitions exiting the considered unmarked state y, by computing each target state $y' = (x', \delta', i')$ based on $Map(\mathcal{X})$. When there is no unmarked state with $|\delta| = \mathcal{C}$, the cardinality $\mathcal{C}$ is incremented (line 17), and the main loop (lines 3–18) is iterated based on the new cardinality. The algorithm stops when all states are marked, in other words, when $Abd(\mathcal{X}, \mathcal{O})$ is complete.

Example 10 Based on $Abd(\mathcal{P}, \mathcal{O})$ in Fig. 6.2, traced in Table 6.3 is the execution of PRIORITIZED DIAGNOSIS ENGINE for $\mathcal{P}$, where $\mathcal{O} = [ov, oz, oz, ov]$.[5] Within the

[3] The input parameter $Space(\mathcal{X})$ is assumed not to be materialized, since in real applications its generation is prohibitive owing to the exploding number of states. Hence, the loop statement in line 7, which considers each transition in $Space(\mathcal{X})$, is only a formal specification that is implemented in practice by considering each transition that is triggerable in state x, the latter being expressed as the vector of component states and the vector of events within links.

[4] As a side effect in constructing the abduction, the algorithm generates *spurious* states also, that is, states that are not connected to any final state, which are therefore irrelevant.

[5] Spurious states are missing in Table 6.3, as they do not contribute to candidates (cf. footnote 4).

Algorithm 1: PRIORITIZED DIAGNOSIS ENGINE

input : $\mathcal{X}$: active system with $Space(\mathcal{X}) = (\Sigma, X, \tau, x_0)$ and mapping table $Map(\mathcal{X})$,
$\mathcal{O} = [o_1, \ldots, o_n]$: a temporal observation of $\mathcal{X}$
output: $Abd(\mathcal{X}, \mathcal{O}) = (\Sigma, Y, \tau', y_0, F)$: the diagnosis abduction of $\mathcal{X}$ based on $\mathcal{O}$,
Δ^*: an ascending ordering of the candidate set $\Delta(\mathcal{O})$ generated incrementally

1 $\Delta^* \leftarrow [\,]$, $\mathcal{C} \leftarrow 0$ `// C denotes the (increasing) cardinality of candidates`
2 Create the unmarked initial state of $Abd(\mathcal{X}, \mathcal{O})$, namely $y_0 = (x_0, \emptyset, 0)$
3 **repeat**
4 **while** *there is an unmarked state* $y = (x, \delta, i)$ *in* Y *where* $|\delta| = \mathcal{C}$ **do**
5 **if** $i = n$ **and** $\delta \notin \Delta^*$ **then**
6 Output δ and append it to Δ^*
7 **foreach** *transition* $\langle x, t, x' \rangle$ *in* τ **do**
8 Let (t, o, f) be the triple in $Map(\mathcal{X})$ relevant to component transition t
9 **if** $o = \varepsilon$ **or** $(i < n$ **and** $o = o_{i+1})$ **then**
10 $i' \leftarrow$ **if** $o \neq \varepsilon$ **then** $i + 1$ **else** i
11 $\delta' \leftarrow$ **if** $f \neq \varepsilon$ **then** $\delta \cup \{f\}$ **else** δ
12 $y' \leftarrow (x', \delta', i')$
13 **if** $y' \notin Y$ **then**
14 Insert y' into Y
15 Insert $\langle y, t, y' \rangle$ into τ'
16 Mark y
17 $\mathcal{C} \leftarrow \mathcal{C} + 1$
18 **until** *all states in* Y *are marked.*

Table 6.3 Tracing of algorithm PRIORITIZED DIAGNOSIS ENGINE (cf. Fig. 6.2)

Cardinality ($\mathcal{C}$)	New states in $Abd(\mathcal{P}, \mathcal{O})$	Candidates generated
0	0, 1	
1	2, 3, 4	
2	5, 6, 7, **8**, **11**	$\{fz_3, fv_3\}$
3	9, **10**, 12, **15**, **18**, 13, 14	$\{fz_3, fz_4, fv_3\}$, $\{fz_3, fv_3, fv_7\}$
4	**16**, **17**, 19	$\{fz_3, fz_4, fv_3, fv_7\}$, $\{fz_3, fv_3, fv_4, fv_7\}$
5	**20**	$\{fz_3, fz_4, fv_3, fv_4, fv_7\}$

loop in lines 4–16, for each cardinality $\mathcal{C} \in [0\,..\,5]$, both the set of new states in $Abd(\mathcal{P}, \mathcal{O})$ and the set of new candidates are shown. Final states are in bold. Boxed (yellow) states are those generated for a given $\mathcal{C}$ but their cardinality is actually $\mathcal{C} + 1$. This comes from the loop in lines 7–15 generating the transition function of an unmarked state y, which amounts to computing *all* the transitions $\langle y, t, y' \rangle$ exiting

y, with $y' = (x', \delta', i')$, where $|\delta'|$ is possibly $\mathcal{C} + 1$ (when t is faulty) rather than $\mathcal{C}$. The candidates associated with these extra states are output at the next iteration, after incrementing $\mathcal{C}$ (lines 17).

6.6 Experimental Results

The diagnosis technique presented in this paper was implemented in the *C* programming language, under the *Linux* operating system (distribution *Kubuntu* 20.04), running on a laptop with 16GB of working memory. The software package allows for the specification of active systems by a specially-designed language. Once the description of an active system is compiled into internal data structures, the relevant space, such as *Space*($\mathcal{P}$) in Fig. 6.1, can be generated as a graph specified in the *dot* language by exploiting the *Graphviz* package.[6] If a temporal observation is given, the relevant abduction can be generated also, like *Abd*($\mathcal{P}$, $\mathcal{O}$) in Fig. 6.2. Two engines have been implemented for the generation of candidates: DIAGNOSIS ENGINE, which outputs the candidates without any prioritization, and PRIORITIZED DIAGNOSIS ENGINE, the algorithm presented in this paper. The two algorithms were compared based on several experiments. Shown in Fig. 6.3 are the results relevant to the abduction displayed in Fig. 6.2 for both DIAGNOSIS ENGINE (left) and PRIORITIZED DIAGNOSIS ENGINE (right). Specifically, in each bar chart, the x-axis indicates the order number of the candidate generated, while the y-axis on the left indicates the cardinality of the candidate. Hence, a bar in position $k \geq 1$ having height h corresponds to the output of the k-th candidate, whose cardinality is h. For instance, the first candidate generated by DIAGNOSIS ENGINE includes three faults, while the first candidate generated by PRIORITIZED DIAGNOSIS ENGINE includes two faults. The additional y-axis on the right indicates the processing time (in μs) spent so far, whose value is represented as a red bullet. For instance, the time spent by DIAGNOSIS ENGINE up to the generation of the fourth candidate, is 22 μs, while the generation of the same (most probable) candidate by PRIORITIZED DIAGNOSIS ENGINE is almost instantaneous. DIAGNOSIS ENGINE completes the generation of the candidate set in 26 μs, while PRIORITIZED DIAGNOSIS ENGINE takes 27 μs. Albeit these figures have little significance owing to the small dimension of the problem, they nevertheless reflect the need for a slight additional computation by PRIORITIZED DIAGNOSIS ENGINE for ranking and searching the candidates. On the other hand, unlike DIAGNOSIS ENGINE which generates an unordered set of candidates, PRIORITIZED DIAGNOSIS ENGINE generates the candidates in ascending order, thereby providing more probable candidates as soon as possible.

To compare more significant figures, another experiment with a different mapping table of $\mathcal{P}$ is presented, which involves ten faults and just one observation label. Given in input a temporal observation of length 60, the diagnosis results (including 205 candidates) are shown in Fig. 6.4, where the processing time is expressed in

[6] *Graphviz* is a shorthand for *Graph Visualization Software*, which is available at *graphviz.com*.

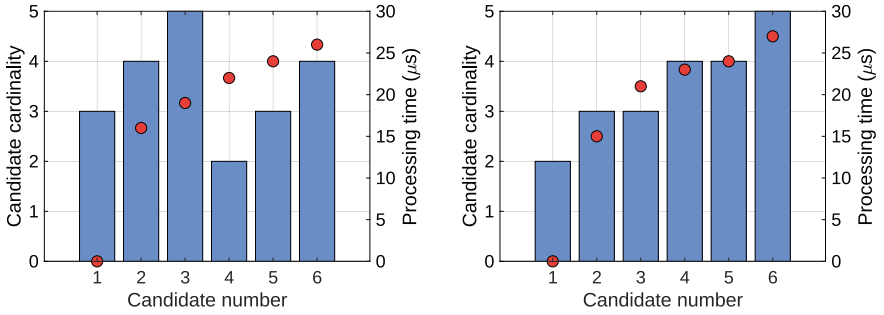


Fig. 6.3 Comparison for $\mathcal{P}$ (cf. $Abd(\mathcal{P}, \mathcal{O})$ in Fig. 6.2): DIAGNOSIS ENGINE (left) and PRIORITIZED DIAGNOSIS ENGINE (right), with processing time expressed in μs

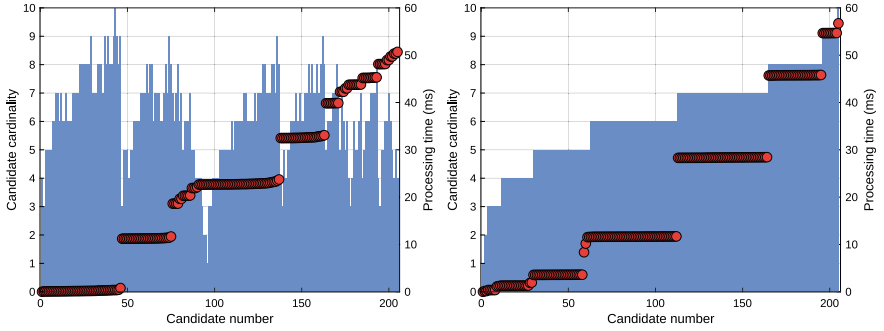


Fig. 6.4 Compared results (including 205 candidates) for a variation of the mapping table and the temporal observation of $\mathcal{P}$: DIAGNOSIS ENGINE (left) and PRIORITIZED DIAGNOSIS ENGINE (right), with processing time expressed in ms

ms. For the generation of the candidate set, DIAGNOSIS ENGINE and PRIORITIZED DIAGNOSIS ENGINE take 50ms and 56ms, respectively, with 55511 states generated in the abduction. The most probable candidate, however, is generated by DIAGNOSIS ENGINE after 22,67ms (in position 96), while it is generated by PRIORITIZED DIAGNOSIS ENGINE almost instantaneously (in first position). Even the two candidates with cardinality 2 (positions 2 and 3) are generated after 0.15ms by PRIORITIZED DIAGNOSIS ENGINE, while they are generated by DIAGNOSIS ENGINE after 22,66ms (positions 94 and 95). In comparison with DIAGNOSIS ENGINE, the generation of the three most probable candidates (one with cardinality 1 and two with cardinality 2) by PRIORITIZED DIAGNOSIS ENGINE allows for saving 99.33% of the time. If we add to the set of most probable candidates those with cardinality 3 and do the math, the time saved is 97.36%.

Results of a third experiment relevant to a different AS composed of five components are outlined in Fig. 6.5, where the model of each component involves two states and six transitions, four of which are spontaneous (that is, not triggered by events in links) and three of which are unobservable, thereby resulting in a diagnosis

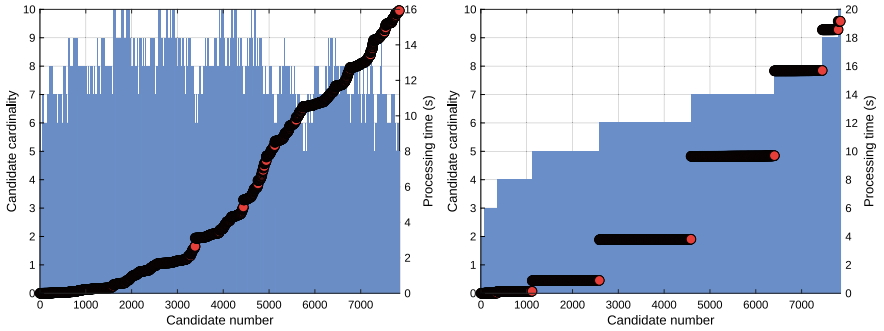


Fig. 6.5 Compared results (including 7848 candidates) for a DES with five components: DIAGNOSIS ENGINE (left) and PRIORITIZED DIAGNOSIS ENGINE (right), with processing time expressed in seconds

abduction with a large number of states (1967090). DIAGNOSIS ENGINE and PRIORITIZED DIAGNOSIS ENGINE took 15.93s and 19.20s, respectively, for generating all candidates (7848), with cardinality ranging from 1 to 10. Focusing on the most likely (low cardinality) candidates, however, figures are telling: the generation of singleton candidates is completed by DIAGNOSIS ENGINE and PRIORITIZED DIAGNOSIS ENGINE in 15.927152s and 0.000694s, respectively, while, considering candidates with one or two faults, the times are 15.930147s and 0.015191s, respectively: compared to DIAGNOSIS ENGINE, the time saved by PRIORITIZED DIAGNOSIS ENGINE for generating the most probable candidates are 99,9956% (one fault) and 99,9046% (one or two faults).

6.7 Conclusion

In the literature on (a posteriori) diagnosis of ASs, the (sound and complete) set of candidates has always been generated without any particular order. Since low-cardinality candidates are more likely than high-cardinality candidates, good heuristics is focusing on the most likely candidates (diagnoses with few faults). This is essential in diagnostic environments that are required to provide the most probable diagnoses as soon as possible so that relevant recovery actions are enabled (possibly automatically) in real time. The PRIORITIZED DIAGNOSIS ENGINE proposed in this paper serves this purpose: candidates are produced in ascending order by cardinality so that most probable diagnoses are generated upfront. Experimental results confirm the effectiveness of the approach. In the future, candidate prioritization may be extended to monitoring-based diagnosis of ASs [9], diagnosis of deep ASs [11], and sequence-oriented diagnosis of ASs [7].

Acknowledgements We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI, under the NRRP MUR program funded by the NextGen-

erationEU, specifically by the project Argumentation for Informed Decisions with Applications to Energy Consumption in Computing – AIDECC (CUP D53C24000530001).

References

1. Abreu, R., van Gemund, A.: A low-cost approximate hitting set algorithm and its application to model-based diagnosis. In: Eighth Symposium on Abstraction, Reformulation, and Approximation (SARA'09), pp. 2–9. AAAI Press (2009)
2. Baroni, P., Lamperti, G., Pogliano, P., Zanella, M.: Diagnosis of large active systems. Artif. Intell **110**(1), 135–183 (1999). https://doi.org/10.1016/S0004-3702(99)00019-3
3. Cassandras, C., Lafortune, S.: Introduction to Discrete Event Systems, 2nd edn. Springer, New York (2008)
4. Grastien, A., Cordier, M., Largouët, C.: Incremental diagnosis of discrete-event systems. In: Sixteenth International Workshop on Principles of Diagnosis (DX 2005), pp. 119–124. Monterey, CA (2005)
5. de Kleer, J.: Hitting set algorithms for model-based diagnosis. In: 22nd International Workshop on Principles of Diagnosis (DX 2011), pp. 60–67. Murnau, Germany (2011)
6. de Kleer, J., Williams, B.: Diagnosing multiple faults. Artif. Intell. **32**(1), 97–130 (1987)
7. Lamperti, G., Trerotola, S., Zanella, M., Zhao, X.: Sequence-oriented diagnosis of discrete-event systems. J. Artif. Intell. Res. **78**, 69–141 (2023). https://doi.org/10.1613/jair.1.14630
8. Lamperti, G., Zanella, M.: Diagnosis of discrete-event systems from uncertain temporal observations. Artif. Intell. **137**(1–2), 91–163 (2002). https://doi.org/10.1016/S0004-3702(02)00123-6
9. Lamperti, G., Zanella, M.: Monitoring of active systems with stratified uncertain observations. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **41**(2), 356–369 (2011). https://doi.org/10.1109/TSMCA.2010.2069096
10. Lamperti, G., Zanella, M., Zhao, X.: Introduction to Diagnosis of Active Systems. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92733-6
11. Lamperti, G., Zanella, M., Zhao, X.: Diagnosis of deep discrete-event systems. J. Artif. Intell. Res. **69**, 1473–1532 (2020). https://doi.org/10.1613/jair.1.12171
12. Pencolé, Y., Cordier, M.: A formal framework for the decentralized diagnosis of large scale discrete event systems and its application to telecommunication networks. Artif. Intell. **164**(1–2), 121–170 (2005)
13. Reiter, R.: A theory of diagnosis from first principles. Artif. Intell. **32**(1), 57–95 (1987)
14. Sampath, M., Sengupta, R., Lafortune, S., Sinnamohideen, K., Teneketzis, D.: Failure diagnosis using discrete-event models. IEEE Trans. Control Syst. Technol. **4**(2), 105–124 (1996)
15. Siddiqi, S.: Computing minimum-cardinality diagnoses by model relaxation. In: Walsh, T. (ed.) Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI 2011), vol. 2, pp. 1087–1092. AAAI Press, Barcelona, Spain (2011)
16. Stern, R., Kalech, M., Feldman, A., Provan, G.: Exploring the duality in conflict-directed model-based diagnosis. In: Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12), pp. 828–834 (2012)
17. Torasso, P., Torta, G.: Computing minimum-cardinality diagnoses using OBDDs. In: Günter, A., Kruse, R., Neumann, B. (eds.) KI-2003: Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol. 2821, pp. 224–238. Springer, Berlin, Heidelberg (2003)
18. Zhao, X., Ouyang, D., Lamperti, G., Tong, X.: Minimal diagnosis and diagnosability of discrete-event systems modeled by automata. Complexity **2020**, 1–17 (2020). https://doi.org/10.1155/2020/4306261

Chapter 7
An Automated Monitoring System for Controlled Greenhouse Horticulture

Matthias Becker and Kinwoon Yeow

Abstract In the field of controlled horticulture, various methods have been studied to facilitate the environmental data retrieval. One of the great findings in this research is the attraction of insect's behaviour towards the LED lighting of various wavelengths. Previous research shows promising results using LED based insect traps for insect population estimation in greenhouses. Therefore, an automated monitoring system is proposed as a standardization tool for environmental data gathering and estimation of pest population in controlled horticulture settings. The proposed automated monitoring system integrates object recognition models (combination of YOLOv3 and SVM) that identify and classify the pest and beneficial population density. The proposed system provides informative output via a mobile application. As a result, the proposed system functions as an integrated IoT management tool that simplifies the information retrieval process.

7.1 Introduction

In the field of agricultural production, many methods have been integrated to facilitate the retrieval process of environmental data. The term "environmental data" in this research includes factors such as humidity, light intensity, temperature, pest and beneficial population density. For temperature and other parameters that can measured with a relative simple single sensor device, the continuous monitoring is state of the art in modern greenhouse settings. However for the monitoring of pest species and beneficial insect populations, no standardized industrial solutions exist.

Pest control in terms of early recognition of possible infestation and rapid counter measures is an important economical aspect in controlled horticulture, since unrecognized infestation lowers the yield, and unnecessary deployment of chemical

M. Becker (✉) · K. Yeow
FG Human Computer Interaction, Leibniz University Hannover, Appelstr. 9A, 30167 Hannover, Germany
e-mail: xmb@sim.uni-hannover.de

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_7

pesticides is expensive. Pest control has been defined as a method used to monitor and limit pest populations in an ecosystem [1]. The effective methods for the control of the population of herbivorous insects are divided into three categories, namely, chemical pest control, natural pest control and biological pest control. The chemical pest control in the agricultural field is widespread and consists of application of pesticide treatments in the greenhouse or open field [2]. Studies show that the targeted herbivorous insects in the long run develop resistance towards the pesticides used. Gradually, more harmful pesticides are developed in order to efficiently kill off the herbivorous insects. Research shows that continuous pesticide treatment may lead to distortion of agricultural systems [3]. On the economical side, chemical treatment can be rather cost intensive, especially when applied preventively, without detailed monitoring and analysis of the current pest species and grade of infestation.

On the other hand, natural pest control is the control towards parameters such as fungus, illumination, wind, temperature, wave and many other environmental parameters to ensure that they are not suitable for pest development. Biological pest control comprises the introduction of predators, parasitoids and/or pathogens in order to control the population of herbivorous insects in a certain environment [2]. In the context of agriculture, predators are insects that actively feed on another species of insects as source of nutrients whereas parasitoids are insects that reproduce by laying their eggs in the body of host which eventually kill the host. Predators and parasitoids used in biological pest control are termed as beneficial insects.

When is the ideal time to induce pest controlling activities? In order to obtain the answer to it, an evaluation of the controlled environment must be performed. Theoretically, the process of gathering greenhouse environmental data is performed in a timely manner as a prevention towards outbreaks of pest infestation. However practically, horticultural systems are not fully technically automated, so that unlike in technical systems, we do not have a continuous global view of all parameters of the systems. Horticultural systems typically involve many manual tasks and include many more uncertainties in the operations. In the subsequent section, the well-known techniques used in environmental data gathering of general greenhouses are tabulated.

7.1.1 Data Acquisition by Manual Inspection

In general, manual visual inspection in data acquisition [4] is the most widely used technique in greenhouses. Previous research shows that sampling by vision techniques is tedious due to the frequent miniature size of individual pest subjects. The process comprises the identification of insect species including different life-stages, the counting of each of those and finally the tallying of all counts. The identification process of top and bottom of a single leaf is as shown in Fig. 7.1. The counts are then recorded accordingly. These data are tabulated for the entire growing season and used as reference for the subsequent seasons as well as for extrapolation of the development of the pest population during the season.

Fig. 7.1 Manual checking on top and bottom of the leaves

7.1.2 Semi-automated Detection with Sticky Trap

For manual visual inspection of the pest infestation, yellow sticky traps are widely used. Also semi-automated approaches have been developed that combine a sticky trap with a camera observing the sticky trap. Several works or products exist in that area. However, there exist drawbacks. The glue of traps will start dripping if the trap is exposed to the temperatures in the greenhouse for a longer time of automated detection. Moreover, the trap might be covered with too many insects, making automated analysis of the photos impossible. In both cases manual interception still is necessary regularly.

The detection of small biological objects such as insects with approximate dimensions of 2mm is a real challenge, especially when considering large commercial greenhouses which are larger than 85 m^2 of area. Hence, it is not possible to perform a continuous daily control and examine every leaf in the greenhouse. Conventionally, visual observations are formed on a weekly basis with coloured sticky traps by human experts. Since this technique does not allow to precisely study the epidemic spatial model, observations on natural support, namely, the manual vision method are favoured. Thereafter, the method of adapted sampling significantly contributes to reduce the amount of data and speed up the analysis process.

In 2018, a semi-automated detection system using the Raspberry Pi with a sticky trap is proposed [5]. The detection has been done automatically using the Raspberry Pi Camera Module v2 with Sony IMX219 8-megapixel sensor, which can be used to take high-definition video, as well as still photographs. Specifically, the model implemented consists of the object detection method based on You Only Look Once (YOLO) [6] and the classification method based on Support Vector Machines (SVM) [7]. The mean classification accuracy is 90.18% and the mean counting accuracy is 92.50% on Raspberry Pi [5]. It is still semi-automated as the usability of sticky trap is required to be manually changed in a regular interval. Hence, the semi-automated detection system with sticky trap is available commercially.

7.1.3 Automated Pheromone-Based Trap

Similarly, the automated pheromone-based detection system is designed to attract specific gender of insect species with their pheromone [8]. The developed detection system is done with RetinaNet which combines two feature extraction algorithms ResNet and FPN for classification and bounding box regression. The research shows a mean object recognition precision of 74.6% with mean run time of 23.44s. Nowadays, there are many companies targeting on improvising this setup and sell them commercially. Therefore, getting the correct tool for data acquisition is important in the field of horticulture.

7.2 Proposed Automated Monitoring System

From the drawbacks of existing semi-automated data collection frameworks, we propose an approach circumventing the drawbacks of sticky traps in automated pest monitoring systems.

7.2.1 Research Background

The research in [9] demonstrates that *T. Vaporariorum* is highly attracted to green LED light. The attraction of *T. Vaporariorum* to monochromatic green light is caused by a wavelength specific behaviour [10]. Like in other herbivores, such as aphids, this behaviour should be based on two different photoreceptors which are sensitive to blue and green light. A competitive interaction between both, called opponent mechanism enables the insects to discriminate targets according to the reflection pattern in the blue, green and yellow range, independent from the reflection intensity [11]. This mechanism explains the aphids's unexpected preference towards yellow compared to green. The visible spectrum of yellow is approximately 580 nm of wavelength and contains higher reflection intensity compared to green and lower reflection in the blue wavelength range. Experiment shows that *T. Vaporariorum* favours the narrow bandwidth green LED as the green sensitive photoreceptor is stimulated.

7.2.2 System Design

The setup of the proposed monitoring system consists of two independent components interacting via an intermediately server. The connections made in and out of the server are recommended to be established through HTTPS Internet Protocol as all appropriate security features are provided. An individual monitoring device consists of two components, namely, the I/O devices and a Raspberry Pi.

The main components of the I/O device are the LED that lights up from the base of the box as the attraction towards insects and a camera that is used to gather surrounding data.

Other subsidiary components such as photoresistor (light-dependent resistor, LDR) and push button are used in order to facilitate the overall usability of the automated monitoring system. In particular, the photoresistor is used as the detection media of surrounding light condition. Once the surrounding light condition is known, the Raspberry Pi sends the required signal in order to make some adjustments toward the brightness condition of the LED base. In addition, the push button is configured in such a way that the Wi-Fi Protected Setup (WPS) is established for each individual monitoring device. Therefore in each use case of the proposed monitoring system, one or more monitoring devices are connected to a single user.

The items used for I/O device in our implementation consist of Raspberry Pi 4 Model B with 4GB RAM, UEye 1007XS-C camera, Sunfounder photoresistor and a push button.

7.2.3 Mobile Application

For acceptance of the automated monitoring system it has to be user-friendly. It should be easily installed and maintained and used by the greenhouse staff. In the user point of view, the developed mobile application is able to execute monitoring tasks as follows:

- The configuration of each individual monitoring device.
- The observation of monitoring data and some simple analysis.
- The flexibility of changing the colour of LEDs.
- The setting of threshold limit that user could obtain notification alert.

The primary task of the whole system is to monitor the overview condition of greenhouse environment. Figure 7.2 shows the mobile interface on setting up the device. In Fig. 7.2, the editing and renaming of individual monitoring device is shown on the left most, while the result of the renaming process is shown in the middle of the figure. In order to facilitate the process of monitoring several devices, the application is able to dynamically add devices with either scanning of QR code on individual monitoring device or entering its device id.

Figure 7.3 shows the configurations available and result of data tabulation on the mobile interface developed. The leftmost interface in Fig. 7.3 illustrates on the selection of light spectrum that is emitted on the LEDs of monitoring device. The LEDs spectrum is crucial in attraction of desired insect species as mentioned in the earlier section. Hence, the user is able to adjust the light spectrum in attracting the appropriate insects for the desired outcome. The middle interface on Fig. 7.3 shows the setting of threshold limit per frame basis on each insect species. These thresholds are subsequently reflected on the tabulated graph. Furthermore, the notification on

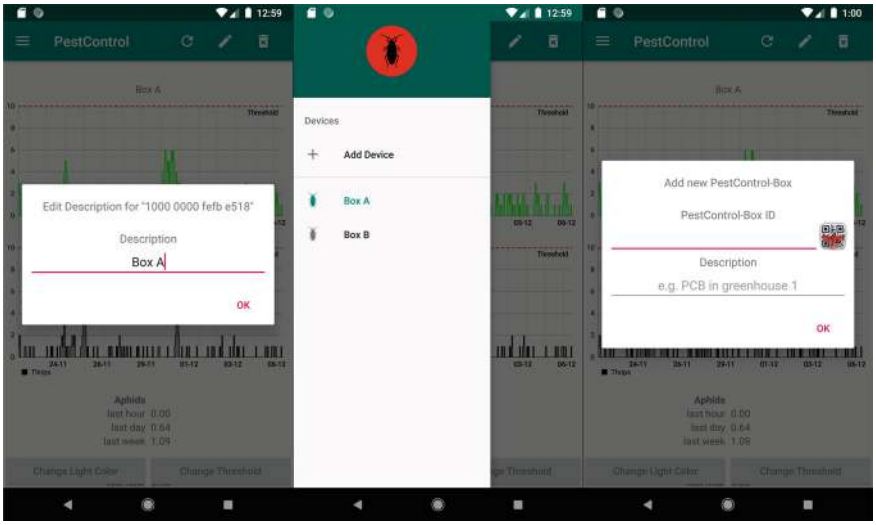


Fig. 7.2 Mobile interfaces shown during setting up devices

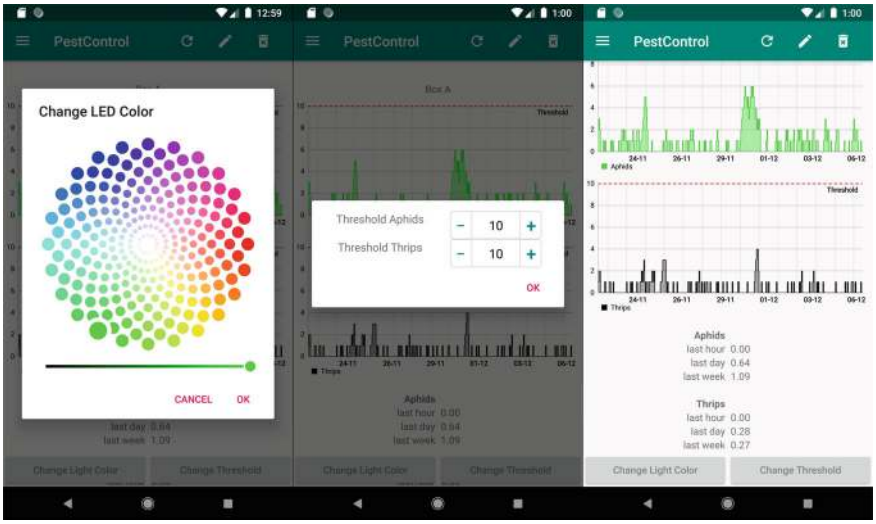


Fig. 7.3 Mobile interface of I/O density threshold and LED wavelength

receiving threshold alert of insect population density is one of the configurations shown. Once the desired thresholds are determined, alert notification will be received when surrounding insect population density reached or surpassed its thresholds. The right most interface shown in Fig. 7.3 indicates the result tabulation in graphical form and a simple analysis of the data collected.

7.2.4 Software Framework

According to our literature, the software framework that achieves the highest accuracy up-to-date of 90.18% is proposed by Zhong et al. and is adopted [5]. The detection and classification models are pre-trained on a local machine. Once the image is obtained from the I/O device, the Raspberry Pi executes the YOLOv3 detection algorithm on the acquired image. Next, the extracted features are used in the SVM classification algorithm. In this software model, the pre-trained models used are crucial in analysis of result. The implementation of this framework model is suitable as execution of models loaded in the computing device, Raspberry Pi is relatively fast. Hence, it is suitable for our purpose of real-time computing.

7.3 Result and Analysis

The developed prototype is shown in Fig. 7.4. The details of each I/O device are as described in Sect. 7.2.2. Figure 7.4 shows the overview of implemented prototype on different perspectives and various lighting conditions. This monitoring device is considered as one element of the entire proposed system. On the bottom left of Fig. 7.4 shows the I/O implanted on the Raspberry Pi, namely, photoresistor that is used in brightness adjustment. Subsequently, a WPS push button is installed on the right of the photoresistor for the simplification of user's Wi-Fi configuration.

7.3.1 Object Recognition Results

Two experimental trials of 14 days each are executed with the implemented prototypes. In each trial, two cages of dimensions (35 cm $\times$ 70 cm $\times$ 35 cm) are treated with pest population of *Aphis Gossypii* and *Frankliniella Occidentalis* respectively. The sample labelled images data gathered on two insect species for training phase are as shown in Fig. 7.5.

The training result for YOLOv3 detection model developed is as shown in Fig. 7.6. The loss functions on both training and validation indicate that only small differences between both loss values are found after approximately ten epochs of iterations. Hence, the training is performed sufficiently to prevent huge training loss.

The result of SVM classification with 10-fold cross-validation from the training phase is tabulated in Table 7.1. The F1-score shows that sufficient training has been performed for the SVM classification.

Figure 7.7 shows the detection process of both pest species with bounding box on each individual object, using the YOLOv3 model developed. Our preliminary study shows that the execution of object recognition including both detection and classification achieves mean accuracy of 83.91% on both classes of pest species.

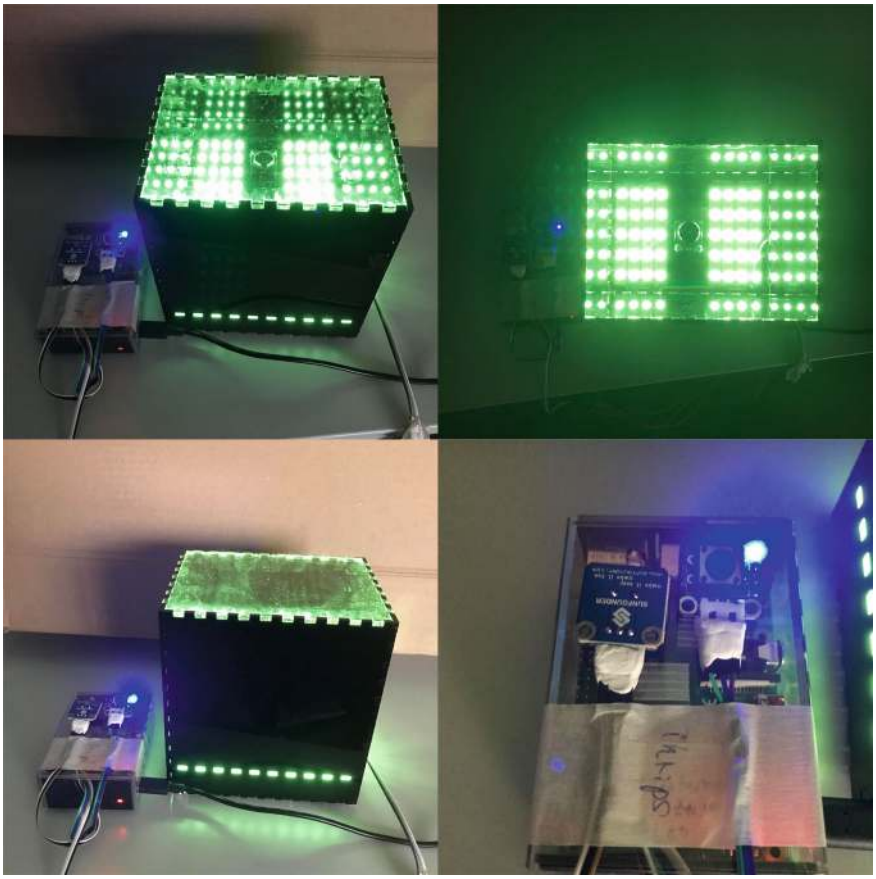


Fig. 7.4 Implemented prototype of individual monitoring device

Fig. 7.5 Labelled data on *Aphis Gossypii* (left) and *Frankliniella Occidentalis* (right)

Fig. 7.6 Loss function of YOLOv3 training

Table 7.1 Result of SVM 10-fold Cross-Validation

Class	Precision	Recall	F1-score	Support
A. Gossypii	0.99	1.00	1.00	481
F. Occidentalis	1.00	1.00	1.00	481
Accuracy			1.00	962
Macro avg	1.00	1.00	1.00	962
Weighted avg	1.00	1.00	1.00	962

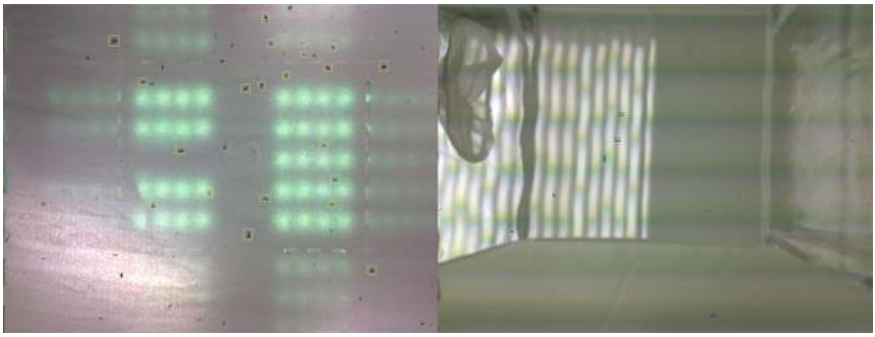


Fig. 7.7 Detection of individual pest object on images

7.4 Conclusions

In our work we present a fully automated insect monitoring system. It is proposed and implemented for the controlled horticulture environment based on secure "wireless network" topology that enables various environmental data to be transferred in real-time. One of the main advantages of a fully automated monitoring system compared to previous semi-automated trap approaches is that semi-automated traps require regular maintenance as traps have to be changed on a regular basis. Contrary to that, our approach, using adjustable LED illuminated surfaces to attract and count samples of the insect populations in the greenhouse, enables the automated trap to work correctly and without maintenance for a long time (optimally one full growing season), while in other approaches, dripping glue and filled sticky traps still necessitate manual intervention of the greenhouse staff. The IoT enhancement developed in the proposed system is one of possible solutions for reducing the coloured sticky traps and pheromone traps waste produced in the activity of monitoring population density. The proposed automated monitoring system aims to reduce the tedious work on data acquisition of population density. In this paper, possibility of controlling the setup of the monitoring device using an Android-based application is presented. The result is generated in a readable manner and tabulated in a graphical form for the users. In the current prototype, the notifications are set on individual thresholds per frame basis. Thereby, the individual threshold could be calibrated into certain duration basis in the future.

The concept of using single Raspberry Pi as a computing unit for object recognition processes is proven to be sufficient. Furthermore, the Raspberry Pi provides flexibility of integrating various I/O devices as tools to gather the environmental data required, namely light intensity, air humidity, air temperature and soil temperature. Our preliminary results concerning object detection and classification success are relatively lower compared to other results in the literature. This might partly be caused by variable lighting conditions including uneven background brightness due to the use of LEDs, and also by the relatively small size of our target insects.

References

1. Wilby, A., Thomas, M.B.: Are the ecological concepts of assembly and function of biodiversity useful frameworks for understanding natural pest control? Agric. Forest Entomol. **4**(4), 237–243 (2002)
2. Marrone, P.G.: "Barriers to adoption of biological control agents and biological pesticides," Integrated pest management, pp. 163–178. Cambridge University Press, Cambridge (2009)
3. Wilson, C., Tisdell, C.: Why farmers continue to use pesticides despite environmental, health and sustainability costs. Ecol. Econ. **39**(3), 449–462 (2001)
4. Boissard, P., Martin, V., Moisan, S.: A cognitive vision approach to early pest detection in greenhouse crops. Comput. Electron. Agric. **62**(2), 81–93 (2008)
5. Zhong, Y., Gao, J., Lei, Q., Zhou, Y.: A vision-based counting and recognition system for flying insects in intelligent agriculture. Sensors **18**(5), 1489 (2018)

6. Jocher, G.: ultralytics/yolov5:v3.1-Bug Fixes and Performance Improvements (2020). https://github.com/ultralytics/yolov5. https://doi.org/10.5281/zenodo.4154370
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
8. Sun, Y., Liu, X., Yuan, M., Ren, L., Wang, J., Chen, Z.: Automatic in-trap pest detection using deep learning for pheromone-based Dendroctonus valens monitoring. Biosyst. Eng. **176**, 140–150 (2018)
9. Stukenberg, N., Gebauer, K., Poehling, H.-M.: Light emitting diode (led)-based trapping of the greenhouse whitefly (trialeurodes vaporariorum). J. Appl. Entomol. **139**(4), 268–279 (2015)
10. Coombe, P.: Wavelength specific behaviour of the whitefly trialewodes vaporariorum (homoptera: Aleyrodidae). J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. **144**(1), 83–90 (1981)
11. Kelber, A., Osorio, D.: From spectral information to animal colour vision: experiments and concepts. Proc. R. Soc. B Biol. Sci. **277**(1688), 1617–1625 (2010)

Chapter 8
Enhancing Music Genre Classification with Artificial Intelligence

Tudor-Constantin Pricop and Adrian Iftene

Abstract In this paper, we investigate a comprehensive approach to music genre classification, utilizing a combination of deep neural networks and machine learning algorithms. Our research aims to advance the understanding of music's impact on our lives and develop methodologies to create diverse and engaging musical experiences tailored to individual preferences. We begin by extracting relevant features from a large and diverse collection of music samples from different genres. These features, encompassing spectral properties, rhythmic patterns, and tonal characteristics, serve as the foundation for our genre classification and generation models. We employ deep neural networks and machine learning algorithms to effectively classify music genres by capturing the distinct characteristics of each genre. While not claiming state-of-the-art performance, our approach demonstrates promising outcomes in classification tasks, showcasing its potential to enhance music-related applications such as recommendation systems and creative tools for composers.

8.1 Introduction

The transformative power of music has been a cornerstone of human culture and expression throughout history. As diverse as our societies, the vast array of music genres reflects our rich cultural heritage and individual creativity. With the rapid advancements in technology, new paradigms in music analysis and synthesis are emerging, further enhancing our understanding of this universal language. The motivation for this research paper stems from a desire to harness the potential of cutting-edge learning techniques, such as Machine Learning (ML) and Deep Neural Networks (DNNs), to address one primary challenge in the field of music informatics: genre

T.-C. Pricop · A. Iftene (✉)
"Alexandru Ioan Cuza" University, Iasi, Romania
e-mail: adiftene@info.uaic.ro; adiftene@gmail.com

T.-C. Pricop
e-mail: tudor.pricop@info.uaic.ro

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_8

classification [1]. The accurate classification of music genres is a crucial task that aids in music discovery, recommendation, and tagging [2]. However, the subjective nature of genre definitions and the complexity of musical elements make this a challenging problem. By employing expertise in music theory, ML, and DNNs techniques, this research aims to develop robust and sophisticated models that can effectively capture the intricate patterns and features of various music genres, leading to improved classification performance. It is hoped that the findings will not only serve as a foundation for future research endeavors but also pave the way for innovative applications that enrich our understanding and appreciation of the world of music.

8.2 Related Work

In paper [3], the authors worked with the GTZAN dataset, using both ML and DL techniques. The features and Mel spectrograms used were computed by the team, and as a dimensional reduction technique, they used Principal Component Analysis. The models used for experimentation: K-Nearest Neighbors, Support Vector Machine, Feed Forward Neural Network, and Convolutional Neural Network (CNN). Following data processing methods and training the models for a large number of epochs, the results showed a maximum of 60% accuracy for ML algorithms and a dramatic jump of performance with Mel spectrograms, at approximately 82%. Regarding the visible overfitting of each model, the authors introduced batch normalization, dropout layers, and L2 regularization.

Looking at the confusion matrices, the genres that were hardest to classify were Rock (due to lack of visible distinctive traits in spectrograms) and Jazz (piano-heavy tracks may have been too close to classical music). In [4], another dataset was used for the task of genre classification, the small subset of the Free Music Archive, which contains 8,000 tracks categorized into 8 top-level genres. They experiment with both features and Mel spectrograms, making use of CNNs, Convolutional Recurrent Neural Networks (RNNs) with 1D and 2D convolutions, and RNNs. To combine the advantages of each model, they also use ensembles, which achieved similar and even better results than previous state-of-the-art methods. As they delve into their work, they also emphasize that the task is very challenging, even for music experts, especially when some genres are vaguely defined and interpolated with other genres (Pop, Experimental).

It is also mentioned how ambiguous the conclusions can be between different research papers that study the same task, as the databases, the genres, and even the measurements vary. A larger study on the GTZAN dataset [5], provides a comparative analysis of genre classification using deep learning and ML techniques, obtaining impressive results compared to literature counterparts. They also extract various features from the audio signals and use information gain ranking to select the most relevant ones, to reduce the amount of irrelevant or redundant data. Regarding practical applications of the above-mentioned methods, one of the most popular

applications is **Spotify**, a digital music, podcast, and video streaming service that gives access to millions of songs and other content from creators all over the world [6]. Artificial Intelligence (AI) plays an integral role in Spotify's functionality, analyzing the acoustic, cultural, and personal inputs for every user, to create personalized recommendations. There are even some AI-driven systems that predict upcoming hits based on users' behavior patterns. Moreover, Spotify is developing several models to create music and provide songwriting assistance, showcasing AI's potential in the creative aspects of the music industry.

8.3 Proposed Solution

8.3.1 Datasets

When seeking the perfect dataset for the music genre classification task, it is essential to consider the size of the dataset, its diversity, balance, and the quality of the genre labels. A synthesis of many musical datasets is presented in [7]. For our experiments, we took into account only three of the most commonly utilized datasets for this purpose: GTZAN [8], Free Music Archive (FMA) [9], and the Million Song Dataset (MSD) [10]. Each dataset has its unique strengths and weaknesses that were carefully analyzed before choosing the perfect candidate for analysis and experimentation. Usually, **GTZAN** is the first choice for music genre recognition tasks, primarily due to its simplicity, reduced size, and balanced structure. It provides a neat, evenly distributed collection of a thousand audio tracks across ten popular genres. Because of the balanced genres, the need for additional preprocessing is eliminated; otherwise, the class imbalance could have caused bias in the models. The collection contains a folder with the 30-second audio files, each subfolder is labeled with one of the 10 genres and contains the respective 100 files. The other folder contains the visual representation of each audio file (Mel Spectrograms), also distributed in 10 balanced subfolders.

The **Free Music Archive (FMA)** dataset is a richly annotated, high-quality, and diverse collection of music that can be utilized for various research tasks related to music analysis, including genre tagging. It provides several levels of annotation, for each track there is a unique ID, title, album, and release year. Additionally, there are track genres and sub-genres, providing a robust groundwork for a broad spectrum of music analysis tasks. Tracks are categorized according to a hierarchical taxonomy that spans 16 top-level genres (like Pop, Rock, and Electronic), which are further broken down into 161 sub-genres (such as Psych-Rock, Lo-Fi, and Drone). One important feature of this dataset is its availability in subsets of different sizes: small, medium, large, and full versions.

The **Million Song Dataset (MSD)** is a publicly accessible collection of audio features and metadata for a million contemporary music tracks. With a million songs included, its sheer size offers an extensive range of data for all kinds of tasks in music

research, including music genre classification. Each song includes data points like the release year, the artist, the popularity, the key, tempo, or duration. In addition to this, there are segments, bars, and beats indicating the rhythm and timbre of audio segments, which give an idea of the sound color and melody. The data for each track has been computed using an API called The Echo Nest, which gives objective information about each song, unaffected by subjective factors like personal opinion or cultural context. However, there are also a few drawbacks to this dataset. Firstly, it includes only metadata and precomputed features, it does not contain the actual audio files for the songs, due to copyright reasons. For researching deep learning techniques using Mel spectrograms, this dataset is limited. Secondly, the MSD is heavily weighted toward popular Western music. Even if it offers an extensive range of data for this segment, it does not provide a diverse or representative sample of all genres within world music.

Additionally, we use the **Spotify API**, which is a rich source of music data, providing access to a wide variety of information such as track details, artist information, album details, playlists, and even audio analysis and features, but the most important thing is that it contains genre labels for each track, making it a suitable source for building up a dataset for music genre recognition. Downloading from the Spotify API requires client credentials obtained after making an account on the Spotify Developer platform. The genres used to quiz the API are the following: blues, classical, country, disco, reggae, metal, hip-hop, jazz, pop, and rock. It can be seen that the list is very similar to the one used in GTZAN, to make pertinent comparisons between the 2 datasets. The final version of the dataset contains two folders, $Audio\_Spotify$ with 500 audio tracks per genre and $Audio\_Spotify\_Test$ with 300 tracks per genre. Regarding Deep Learning methodologies, for each audio track, 4 Mel spectrograms were created of shape (256, 256): the original, unmodified version, one with a random frequency mask, one with a time mask, and the last one which has noise added to it (noise with mean 0 and standard deviation 0.3).

This process is called Data Augmentation and it is used to increase model robustness and to avoid overfitting in neural networks [11]. Thus, the dataset of visual representations reached a total of 32,000 files, distributed evenly in 10 folders. For the ML techniques, 2 CSV files were created, found in the original dataset, that contain the mean and variance of each song for several features, including zero-crossing rate, spectral features, harmonics, percussive, mfcc, and others. These audio properties were computed, similarly to the GTZAN dataset, for the entire 30-second tracks and smaller 3-second windows (to increase the amount of data fed to the models). The custom-made **Spotify** dataset allows us to curate audio samples and tailor them to this research specifics, leading to improved model performance and better contrastive results. This vast music library provides a diverse range of genres, artists, and tracks, keeping the research continually updated to reflect current music trends and emerging genres. However, significant effort must be put into cleaning the dataset, extracting features, handling missing values, and ensuring consistent labeling. Furthermore, Spotify's music catalog might induce bias due to popular tracks dominating the dataset, potentially affecting model generalization.

8.3.2 Machine Learning

One important step to do before the feature selection algorithm and also before splitting the dataset into train and test data is normalization. The final choice was to use MinMaxScaler, which scales and transforms the features into the range [0,1], to keep the values on a positive note because the distribution of data is unknown. While working with the 2 datasets, it has been observed that the models work better with more data, even if the small 3-second window is not big enough to encapsulate the full spectrum of changes that can appear in a song. When looking at the feature selection step, the first thing to do is to apply filter methods to understand the structure of the data and the impact of individual variables. Each feature is ranked according to a score that reflects the feature's relevance in predicting the target variable. The most relevant features were $flux\_var$, $percussive\_var$, $spectral\_centriod\_var$, rms_mean and rms_var. Then, wrapper and embedded methods are applied to a series of ML algorithms to find the best combination of features that contribute to the accuracy of the model. These methods are integrated into the pipeline that is used for model training, evaluation, and hyperparameter tuning (Fig. 8.1).

For all experiments, five-fold cross-validation was used on the training data. This was combined with a grid search to help find the best-performing hyperparameters (Table 8.1).

Cross-validation allows the classifier to train on more data than it would need if both validation and test set had to be used. Grid search ensures that each combination of parameters is chosen and evaluated and eventually reaches the optimum set, the one with the maximum accuracy score. Following this *interface*, several algorithms were trained: KNN, SVM, Random Forest, Logistic Regression, Naïve Bayes, and XGBoost. Each was evaluated on the test set and a confusion matrix was made to visualize the model's predictions compared to the actual labels of the dataset and to assess which genres are easily recognized and which are similar. In the end, multiple models were combined into an ensemble model, to improve accuracy, generalize

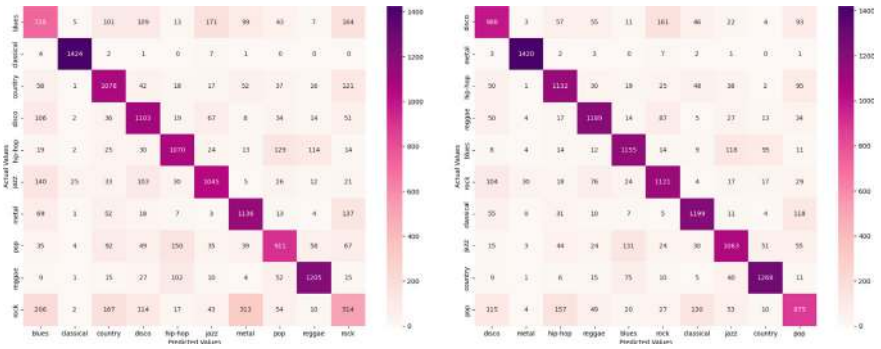


Fig. 8.1 Confusion matrix of the SVM model (left) and the ensemble model, which stacks all the other models (right)

Table 8.1 Example of a classification report-SVM

Gender	Precision	Recall	F1-score
Blues	0.47	0.51	0.49
Classical	0.97	0.97	0.97
Country	0.69	0.76	0.72
Disco	0.69	0.68	0.69
Blues	0.47	0.51	0.49
Reggae	0.86	0.82	0.84
Rock	0.33	0.24	0.28
Accuracy			0.67
Macro avg	0.67	0.67	0.67

better, and offset the weaknesses of some models with the strengths of others. The models are put together using the voting technique softly, meaning that the average predicted probabilities for each class give the output. This final model leads to an improvement in accuracy and robustness over the individual models.

8.3.3 Deep Learning

In this section, various DL algorithms are explored and their potential to enhance the model's predictive power and adaptability is examined. Given the complexity and richness of music signals, one popular choice of data representation across many research studies is to use Mel Spectrograms. When working with images, there are typically two types of color representations used in neural networks: grayscale and RGB. In grayscale (monochrome), darker shades represent lower amplitudes, while brighter shades indicate higher amplitudes. On the other hand, RGB uses 3 channels and has the potential to convey more information, each color channel can represent different aspects or layers of the audio signals. Both interpretations were used for the genre classification task and there were no visible differences in the models' performances, thus the grayscale was used to speed up further experiments. After reading the Mel Spectrograms from the dataset, the final shape of these images is (256, 256, 1).

Scaling data is a fundamental step in preparing the input for AI algorithms, including when using Mel Spectrograms for DL techniques. Before feeding the input to the networks, a new DataFrame is built using scaled images and then it is randomly split into train and test subsets (30% test partition). After this split, each train and test array must be converted to TensorFlow tensors to be passed as inputs to the framework's neural networks. It also provides on-the-fly data augmentation, applying random rotations, shifts, flips, and zooms to the images as they are loaded into memory. CNNs have been demonstrated to be particularly effective for tasks involving image

analysis, thus their use in processing Mel Spectrograms for genre recognition is a natural fit. CNNs are capable of learning hierarchical representations of the input data, from low-level features such as edges and textures to more complex, abstract patterns. In the context of spectrograms, these could correspond to different musical elements such as rhythm, melody, or timbre. These networks also have a property called translation invariance, which is crucial with Mel spectrograms, as the same musical pattern can be learned and identified from it, regardless of its position in time.

The network used for the classification task is made up of 3 convolutional layers with 32 or 64 filters and a kernel size of 3×3. As kernel initializer, Kaiming ($he\_uniform$) is used, because in some cases the default Glorot initializer can lead to inconclusive results [12]. To prevent overfitting, which appeared before the first 10 epochs in the training process, the Dropout technique was used, which disables a chosen percentage (25%) of the neurons during training, making a more robust model. One of the most popular activation functions used in the hidden layers is the Rectified Linear Unit (ReLU), however, the choice was to use the Exponential Linear Unit (ELU) function, to avoid the dead neurons and vanishing gradients which can appear when working with ReLU. In the end, several callbacks were used to monitor the evolution of the training loop. To prevent wasting computational resources and to stop excessive model training, which may lead to overfitting, EarlyStopping is used. Also, ReduceLROnPlateau is added to improve the model's learning efficiency: when the learning curve reaches a plateau, failing to converge, the learning rate is further decreased. Because many models perform better at a step before reaching the final number of epochs, ModelCheckpoint saves the weights of each epoch for later use of the model. It has been observed that 15 epochs are enough to reach either a convergence or the point at which the model starts to overfit (training loss tends to become 0 while testing loss increases).

Another approach for genre recognition is to combine the strengths of CNNs and RNNs. While CNNs extract features and patterns from audio signals, RNNs take these high-level concepts and understand their evolution over time. This model is very similar to the CNN, with an LSTM cell of 128 units added after the flattened input, this cell being capable of learning long-term dependencies. For each model, the loss and accuracy of each epoch were plotted to understand better how the learning process evolves. The experiments were conducted on the spectrograms of both low quality (256×256) and high quality (512×512), however, the results did not improve, only the computational and temporal resources increased, leading to the conclusion that a better resolution does not help in musical patterns recognition (Table 8.2).

8.4 Evaluation

The subtleties between genres seem to still pose a great challenge when it comes to correctly separating them, showing that even if a song mostly encapsulates one genre according to the musical theory, it can still show signs of playing sequences

Table 8.2 ML algorithms scores

	GTZAN params and accuracy		Spotify params and accuracy	
KNN	n_neighbors:1	92%	n_neighbors:1	77%
SVM	C:100, gamma:1, kernel:rbf	91%	C:10, gamma:1, kernel:rbf	71%
Random	max_depth:None	87%	max_depth:None	65%
Forest	min_samples_split:1 n_estimator:200		min_samples_split:1 n_estimator:250	
Logistic regression	C:0.0001,solver:sag, penalty:none	72%		59%
Naive Bayes	var_smoothing:1e-06	50%	var_smoothing:1e-06	45%
XGBoost	learning_rate:0.1 max_depth:7 n_estimators:200	87%	learning_rate:0.1 max_depth:7 n_estimators:250	67%
Voting ensemble	Default	88%	default	67%
Stacking ensemble	Default	92%	default	79%

Table 8.3 DL algorithms scores

	GTZAN train epochs and accuracy		Spotify train epochs and accuracy	
5-block Dense	50	89%	50	70%
3-block CNN	100	70%	15	56%
ResNetV2	30	70%	–	–
MobileNetV2	100	76%	25	59%
CNN+RNN	–	–	10	50%

that resemble other genres. The performance varies with each classification algorithm and shows a clear limitation when it comes to the visual interpretation of the sound. Regarding ML techniques, Table 8.3 shows the best parameters chosen after running a Grid Search algorithm with 5-cross-validation. For each model, the average accuracy across genres is printed. It seems that the lowest score is acquired by the simple Naive Bayes, while the maximum score is given to a distance-based algorithm like KNN. For this model, the choice of the neighbors hyperparameter for the best accuracy is 1, however, it is a sensitive choice as the decision boundaries become unstable, leading to poor generalization on new data. Even if the hyperparameters are alike, there is a clear gap between the accuracy interval resulting in the GTZAN dataset (50 - 92%) and the Spotify-generated dataset (45 - 77%), with the models keeping the same trend (the rankings of each model tend to be the same). One possible motivation behind this difference may be the fact that the GTZAN dataset is biased toward genre classification [13].

Perhaps another reason could be the weakness of the Spotify API in what concerns the resulting list of songs after a genre-specific query. Further experiments showed another interesting result: a Stacking Classifier with all the low-level models and

a Logistic Regression as meta-classifier has a great potential in separating genres, reaching a maximum accuracy score on both datasets: 92% on GTZAN and 79% on Spotify. Across all experiments with the Spotify-generated dataset, it can easily be seen that *rock* is the most misclassified genre among all the classes, being predominantly confused with *metal*. One main cause might be the fact that there is a considerable amount of overlap and fluidity between the 2 genres. Additionally, there are a lot of sub-genres (*hard rock, heavy metal, progressive rock*) where the boundaries become even blurrier. The second most misclassified genre is *blues*, as it is known to share certain features with other genres like *jazz, rock, and country*, just as is shown in the confusion matrices. Deep Learning experiments start with a comparison between the ML algorithms and a simple neural network. According to the model's results after 50 training epochs, it reaches 89% accuracy on the GTZAN dataset, which is better than 5 of the ML techniques tried before, and 70% accuracy on the Spotify dataset, with the same conclusion. Both models tend to slightly overfit and reach an accuracy plateau within the first 20 epochs (Fig. 8.2).

As previously mentioned, the scores tend to look better for the GTZAN dataset, even in the case of the image representations. The best-performing DL algorithm is the one used to classify the features from CSV files, which results in a lack of critical genre properties in the Mel spectrogram. The overtrained CNN network has an accuracy equivalent to the ResNet and is close to the MobileNet, which leads to the conclusion that it can be further improved to achieve better results. Due to the small amount of data, all these networks seem to be trained quite fast (no more than 130 seconds per step, even in the case of transfer learning). Even if the scores decrease in the case of the Spotify dataset, the results are much better than the random choice (10%).

The best model remains the one that classifies based on features, followed by transfer learning (MobileNet). The combined network of CNN and RNN remains around the same score as the Convolutional Network, even if it has some room for improvement. Unfortunately, better models like Vision Transformers and VGG16 cannot be taken into account as the memory and the time (9 hours/epoch) needed

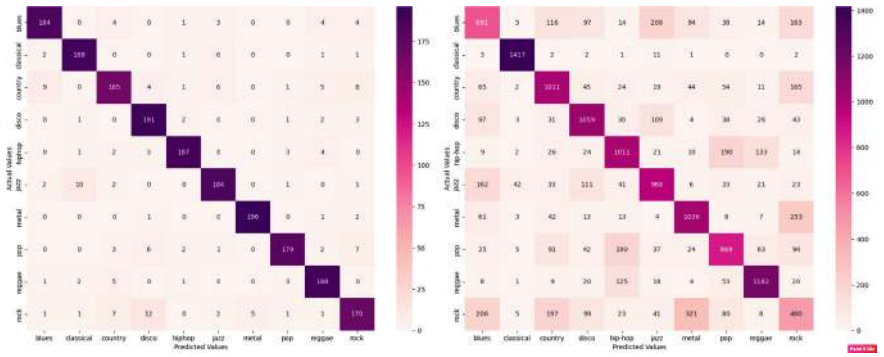


Fig. 8.2 (left) KNN confusion matrix with maximum accuracy on GTZAN, (right) XGBoost confusion matrix on Spotify

for training are inaccessible. Rather than being uniquely integrated into one genre, a single piece of music often transcends boundaries, encapsulating elements from several genres. Consequently, assigning a singular genre to a song is not always accurate, nor does it encapsulate the full spectrum of its musicality. That is the reason why the model will output several percentages showing how many elements from each genre one song contains. The process involves separating the audio track into windows of 3 seconds each, from which several features are extracted and genre membership percentages are computed and summed. There are instances in which the classification models demonstrate remarkable precision, successfully categorizing musical genres with high accuracy, as shown when passing *Alla Turca - Mozart* through the model and receiving great results on the correct label. However, there are also highlighted instances where the models face confusion, struggling to distinguish and correctly classify genres (for example, *Natural–Imagine Dragons*).

8.5 Conclusions

Music genre classification is a complex task that has significant implications in the field of music. The work presented in this paper delved into the intricacies of this task, exploring a variety of advanced ML and DL techniques to classify music genres, based on audio features and Mel spectrograms. A custom-made dataset from Spotify was utilized, as well as the widely recognized GTZAN, to train, evaluate, and compare the models. Each model showcased both strengths and weaknesses. One of the major challenges in music genre classification is the quality and diversity of the data, as music genres are inherently subjective and can overlap, leading to ambiguity and inconsistency in the labels.

Future approaches should deal with classification into multiple classes for a single song, to reflect its musical complexity and richness, acknowledging the fact that a song can belong to multiple genres to varying degrees, thus providing a more nuanced and comprehensive representation of its genre characteristics.

References

1. Lerch, A.: Music Similarity Detection and Music Genre Classification. Wiley (2022)
2. Zhao, Z., Wang, X., Xiang, Q., Sarroff, A., Li, Z., Wang, Y.: Large-scale music tag recommendation with explicit multiple attributes. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 401–410 (2010)
3. Huang, D.A., Serafini, A.A., Pugh, E.J.: Music genre classification. In: CS229 Stanford, pp. 1–6 (2018)
4. Kostrzewa, D., Kaminski, P., Brzeski, R.: Music genre classification: looking for the perfect network. In: ICCS 2021, LNTCS, vol. 12742, pp. 55–67 (2021)
5. Ndou, N., Ajoodha, R., Jadhav, A.: Music genre classification: a review of deep-learning and traditional machine-learning approaches. In: 2021 IEEE IEMTRONICS, pp. 1–6 (2021)
6. Spotify for Developers. https://developer.spotify.com/

7. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: FMA: A Dataset for Music Analysis (2017). arXiv:1612.01840v3. https://arxiv.org/pdf/1612.01840.pdf
8. GTZAN Dataset-Music Genre Classification. https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
9. FMA: A Dataset For Music Analysis. https://github.com/mdeff/fma
10. Million Song Dataset. http://millionsongdataset.com/pages/getting-dataset/
11. Zhang, A., Zhang, H.: Data augmentation techniques for classification of music genre. Appl. Comput. Eng. **33**, 149–156 (2024)
12. Hubens, N.: Why default CNN are broken in Keras and how to fix them. A deep dive into CNN initialization. In: Towards Data Science (2019)
13. Sturm, B.L.: The GTZAN dataset: its contents, its faults, their effects on evaluation, and its future use (2013). https://arxiv.org/pdf/1306.1461

Chapter 9
Quick Image Style Transfer with Convolutional Neural Networks

Bogdan-Antonio Cretu and Adrian Iftene

Abstract In this paper, we are investigating the benefits of employing pre-trained models in computer vision as we work toward creating a model for rapid image style transfer. These models, like VGG or MobileNet, have undergone painstaking optimization for efficiency and precision in picture categorization tasks. We want to change photographs quickly while ensuring high-quality outcomes by utilizing their effective designs and transferable features. The training time needed to create a style transfer model from scratch is considerably reduced by the ability to reuse layers from existing models trained with high computational power on big datasets. We are actively adjusting the models to further tailor them to particular style transfer requirements and aesthetic preferences, allowing for more flexibility and creative control. The methodology we use is based on extracting layers of the pre-trained models with Imagenet dataset weights to create smaller style transfer-capable systems that can be tuned and applied in less than a minute on a particular pair of content and style images. The availability of pre-trained models within well-known deep-learning frameworks, such as TensorFlow and PyTorch, speeds up the implementation of style transfer capabilities and makes the development process easier. Although we are aware of the drawbacks of pre-trained models, we are committed to overcoming them and developing an effective style transfer solution.

9.1 Introduction

In the field of computer vision and image processing known as "style transfer", the style of one image is combined with the content of another. Using this method, it is possible to transform commonplace images into artistic compositions by giving them the distinctive visual elements present in well-known paintings or photographs (see

B.-A. Cretu · A. Iftene (✉)
Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi, Iasi, Romania
e-mail: adiftene@info.uaic.ro

B.-A. Cretu
e-mail: bogdan.cretu@info.uaic.ro

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_9

Fig. 9.1 Example of a content image, style image, and the generated image [1]

Fig. 9.1). Due to its many applications and capacity to produce visually appealing results, style transfer has attracted a lot of attention. Artistic expression is one of the main applications of style transfer. The use of style transfer tools by artists and producers can open up new visual communication possibilities. Artists can experiment with various visual aesthetics, develop original compositions, and arouse particular emotions in their audience by incorporating well-known paintings or artistic styles into their works of art or photography. Due to this, they can develop works that are both visually attractive and compelling while also pushing the limits of their inventiveness. While style transfer has intriguing potential, it has drawbacks. The difficulty of striking a balance between style preservation and content accuracy is one restriction. The perceptual quality of the copied image may occasionally suffer as a result of distortions or changes introduced during the style transfer process. Research is still focused on finding the ideal balance between maintaining the original content's essence and its aesthetic. In what follows, after the presentation of the existing work, we will see the methodologies used for the style transfer on images, followed by the results obtained in the experiments we have carried out.

9.2 Existing Work

Image style transfer has emerged as an important topic in computer vision and graphics. It involves changing an image's appearance to correspond to the features of another image, usually a piece of art. The article [2] separates and recombines image content and style to produce high-quality photos using image representations developed from Convolutional Neural Networks (CNNs) tuned for object detection. Using different layers from the network can be used to manipulate the level of mixing between the content of the initial image and the texture of the style image, [3] provides a deep-learning method for photorealistic style transfer. This method effectively reduces distortion and produces high-quality photorealistic style transfers in a variety of conditions, including adjustments for the season, time of day, weather, and artistic edits results were compared to CNNMRF [4] and Neural

Style [2]. A content- and style-aware stylization strategy for image processing is presented in [5], it uses modules in the model to execute style transformation. Wang et al. [6] examine how well neural style transfer techniques hold up when used with various network designs. To solve biased content representation brought on by locality in CNNs, [7] offers a transformer-based method for image style transfer dubbed StyTr2, which considers long-range dependencies of input images. StyTr2 surpasses state-of-the-art CNN-based and flow-based techniques in qualitative and quantitative studies, proving its usefulness. In the paper [8], a brand-new strategy for learning style feature representations is proposed for transferring arbitrary image styles via contrastive learning: Contrastive Arbitrary Style Transfer (CAST). The paper [9] suggests ArtFlow as a remedy for the content leak issue with universal style transfer. Existing techniques are unaware of the content leak, which occurs when an image's content deteriorates as a result of repeated stylization. To stop content leaks, ArtFlow employs reversible neural flows, an unbiased feature transfer module, and a projection-transfer-reversion algorithm. The limitations of current feed-forward systems are addressed by the method for universal style transfer that is suggested in [10]. To directly match the feature covariance of the content image to a specified style image, the method leverages feature transforms, whitening, and coloring (WCTs), incorporated in an image reconstruction network. The text-driven image style transfer (TxST) that is suggested in this study can transfer any style to images by using written descriptions rather than examples [11]. The authors extract style descriptors from the image-text model (CLIP) and align stylization with the text using powerful image-text encoders and a contrastive training technique. With the aid of the new style transfer technique InST, industrial designers can quickly produce visually appealing products [12]. It creates a neural warping field and texture transformation network from a source product, target product, and style image. InST consists of large-scale geometric warping and interest-consistency texture transfer with mask smoothness regularization terms. It excels at visual product design tasks such as creating logos and bottles. The paper [13] introduces DualStyleGAN, a model that allows for flexible control of dual styles and exemplar-based high-resolution portrait style transfer. Unlike StyleGAN, the model uses both a new extrinsic style path and an intrinsic style path to characterize the content and style of a portrait. The extrinsic style path enables the precise reapplication of the style example through color modulation and complex structural styles. The model's generative space is converted to the target domain using a progressive fine-tuning scheme. According to experimental findings, DualStyleGAN outperforms other state-of-the-art techniques in both flexible style control and high-quality portrait style transfer.

9.3 Methodology for Style Transfer on Images

To evaluate the duration and quality of generated images pre-trained models are used with ImageNet weights [14], this allows leveraging the capacity of existing architectures with lower hardware requirements. VGG19 [15] and MobileNet [16] are two well-known pre-trained models that are used as a baseline for feature extraction.

Fig. 9.2 Content images (left) and Style image samples (right), for each artist. Top to bottom: Alexej Von Jawlensky, Amadeo De Souza Cardoso, Francisco De Goya, Giovanni Boldini, Ohara Koson, Vincent Van Gogh

MobileNet is renowned for its effectiveness and appropriateness for mobile devices, being optimized for quick processing on limited hardware power [17, 18], VGG19 is well known for its success in picture classification tasks [19, 20]. By using both models, we evaluate how well they transferred styles and contrast their advantages and disadvantages, by using different layers we can see the impact of different architectures on the time to generate and the quality of the results. Four different content images were arbitrarily chosen with different structures, a set of dogs, a cityscape, a forest scene, and a desert scene, this allows for a large spectrum of images for the models, and the images are openly available online. For style images, the website Artvee[1] was used, it contains samples of artworks of a large number of artists. For the scope of this work, 6 artists and 6 sample images for each artist were used (see Fig. 9.2).

To compute the loss while training the model we have two components to account for, style loss and content loss. For content, the loss is the difference between the output of the content sublayers when they are applied to the generated image and the original content image. When the newly generated image passes through the content sublayers, the result should be more similar to the outputs from the original image. For style loss the gram matrix [21] can be used to calculate a correlation between the features of the image. These feature correlations capture the texture of the image at different scales, depending on the depth of the layer in the network. For the experiments in this work, the style and content loss were unified in one function applied to the layers marked as content or style within the original model.

[1] Artvee Artists section: https://artvee.com/artists/.

9.4 Experimental Setup

The experiments involve training the model over multiple epochs with specified style and content weights. The optimization process aims to iteratively refine the input image to capture the style of the style image while preserving the content of the content image. The number of steps the model iterates for a pair of content and style is set to 1,000, split into 10 epochs. The experimental configurations can be found in Table 9.1. For each experiment, the parameters are set within a single script that is executed with CLI parameters for each pair of style-content images. The results are saved in a corresponding experiment directory. Each experiment generates all the pairs between the style samples and content images, totaling 144 result images for each experiment. The run time is between 2 and 8 hours for a single experiment, depending on architecture and parameters.

Table 9.1 VGG and Mobilenet experiments: The VGG-Baseline is the example used by Tensorflow to exemplify the usage of pre-trained model layers for style transfer, VGG-1 and VGG-2 exemplify how the quality is affected by different model configurations. The MobileNet experiments are trials of using a smaller model because finding a configuration that blends images with satisfying quality is more difficult

Information experiment	Seconds per generation	Style and Content layers	Image resolution	Total images generated iterations	Style/ Content weights	Optimizer learning rate
VGG-Baseline	178–188	content layers:'block5_conv2' style layers:'block1_conv1', 'block2_conv1', 'block3_conv1' 'block4_conv1', 'block5_conv1'	448×448	144	1e-2/1e4	Adam–0.02
VGG-1	45–49	Content layers: 'block5_conv2' style layers: 'block4_conv1' 'block5_conv1'				Adam–0.02
VGG-2	104–107	Content layers:'block5_conv2' style layers: 'block1_conv1'				Adam–0.02
MobileNet-1	55–58	Content layers:'conv_dw_2_relu' 'conv_pw_3_relu' style layers: 'conv_pw_11_relu' 'conv_pw_13_bn, 'conv_pw_13_relu'				Adam–0.01
MobileNet-2	70–73	Content layers: 'conv_pw_11', 'conv_pw_12, style layers: 'conv_pad_2', 'conv_dw_2', 'conv_pw_2', 'conv_pw_3', 'conv_pw_4_bn', 'conv_pw_5'				Adam–0.03

The experiments were run on a personal computer with the following components: (1) CPU: **Intel Core 10400F** with 32 GB RAM; (2) GPU: **AMD Radeon 6700XT** 12 GB VRAM, NVME storage for images.

9.5 Results

The experiment batches used for presenting the relative performance of MobileNet and VGG19-based models are run with mostly the same parameters, the learning rate and model structure being the differences Table 9.1. The VGG-baseline entry is a basic variant of style transfer that offers great quality by using a subset of layers from VGG19, however, the method takes on average over 3 minutes to generate a single image, this is why other subsets of layers and even MobileNet-based models may reach satisfying results with faster times. Other model configurations were briefly tested for both VGG and MobileNet, but initial results did not point to any qualitative or speed improvement, so the full batches of experiments were not executed.

We can see that in the MobileNet-based experiment batches the content images are suffering very little modifications Fig. 9.3, with only a minor amount of noise being introduced in the image, without specific style properties, this comportment

Fig. 9.3 Examplification of resulted images for three artworks of Alexej Von Jawlensky on the dogs' content image. **a**–MobileNet-1, **b**–MobileNet-2, **c**–VGG-1, **d**–VGG-2

extends to the entirety of the resulted images within all the runs based on MobileNet, some other architectures were tested, without creating an experimental results batch, because no visible improvement occurred with changes in the structure of the style and content models. The VGG-based results have different qualities depending on the layers used, the C - VGG-1 and D - VGG-2 experiments differ by only a few layers, both models utilize a small number of layers from the VGG19 image recognition networks, but their outputs are easily identified after seeing even a single experiment result for each model Fig. 9.3. Going to only two layers, the style detail is lost, but the content image is still modified to a greater degree than it was with the MobileNet-based models. The reason for the big difference in the output of the two VGG-based models is twofold: One, the VGG-1 model utilizes two layers for calculating the style of the image, this allows for information to be better generalized in the model while also losing less information than the single style layer in the VGG-2 experiment; Second, the layers used for extracting styles in the two models are coming from different depths in the model, this is also the cause of the longer execution time of the VGG-2 experiments, although the model has a smaller number of layers chosen for optimization. The position of the layers far from each other in the network causes the propagation of data to be slower and a high loss of detail.

The VGG-1 model has managed to create images with both style and content well preserved Fig. 9.4. The dog content images, line 1, has the strongest content preservation, the original image can be deduced in high proportion by watching the generated result, however going past the main subject and into the less detailed background, the style is mixed to a higher degree with the content, generating a pleasing stylization that is stronger on the background and milder in the foreground. For the images of the trees, line 2, the effect is opposite of the dogs' images, the central point of the image, a bit of space between trees, is most affected by the style transfer, while the rest of the foliage is blended with the style image, causing once again a gradual stylization effect, this time with the central part of the generated image being more textured and having a stronger stylistic effect and the parts of the image closer to the margins being softer and losing some detail from the original content. The trend of separation between elements that belong to the foreground and background continues in the cityscape image, line 3, Larger buildings in the foreground better mix with the style content, and the less detailed smaller buildings in the background blend while the style is applied, making that part of the photo less contrasted. An interesting effect occurs due to the positioning of foreground and background buildings, most foreground ones being on the right side of the content image. The sky in the right sky gets stronger contrast, suggesting that it is considered a part of the foreground, with the bigger buildings, this suggests that a model with a better capacity of separating objects would be able to apply style with more accuracy. The desert dune images, line 4, cause yet another effect to occur. The image has stronger stylistic effects applied around the center of the image, this is likely due to the foreground dune being partially interpreted as a background component The effect is particularly pronounced for image 4B where we can see strong contrast both in the upper multi-dune section of the image, but also in the lower right section where there are footprints on the dunes.

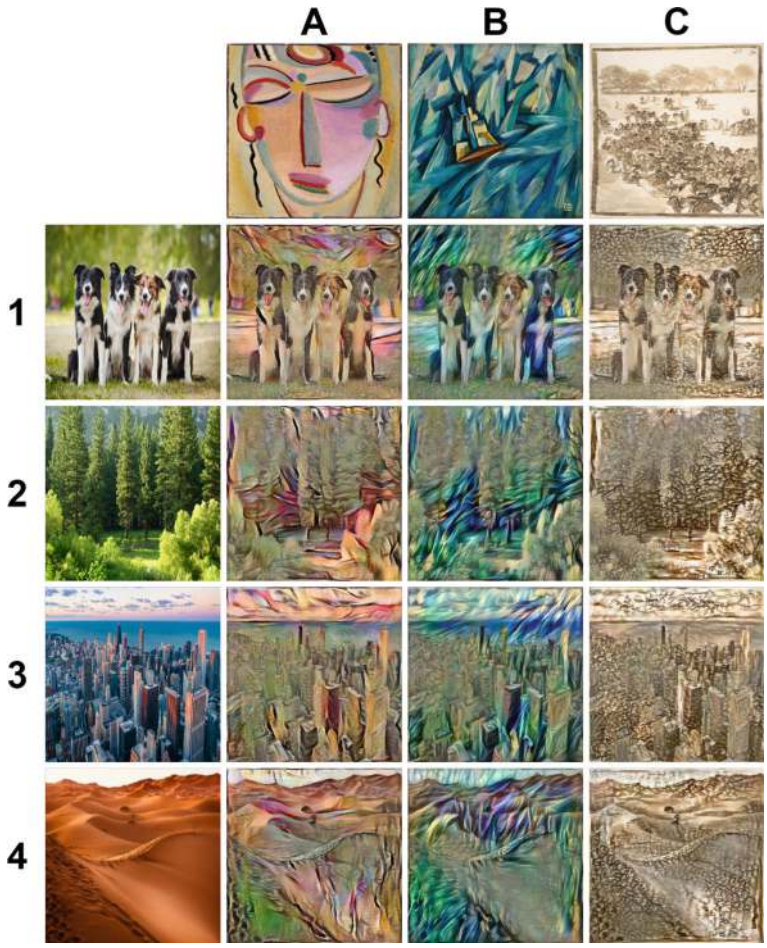


Fig. 9.4 Resulting images of VGG-1 model on all content images and 3 artists, left to right: Alexej Von Jawlensky, Amadeo De Souza Cardoso, Francisco De Goya

The 3-layer-based VGG-1 experiment performed qualitative results, and it is the only one comparable to the baseline approach with 6 layers. Looking at the full images the results of the two models are comparable qualitatively Fig. 9.5.

Watching the images generated for the dogs' image, the biggest difference in generation is the light above the brown dog Fig. 9.5-1, which is smaller on the baseline model. This can be attributed to the multiple layers in the model softening the content of the source, it can be seen as both an improvement or detriment, depending on the target result of the generation. In the generated images for the city content, differences in the two generated images are not immediately apparent. The biggest effect on the quality is the noise that appears in both models, but it is more pronounced in the VGG-1 model.

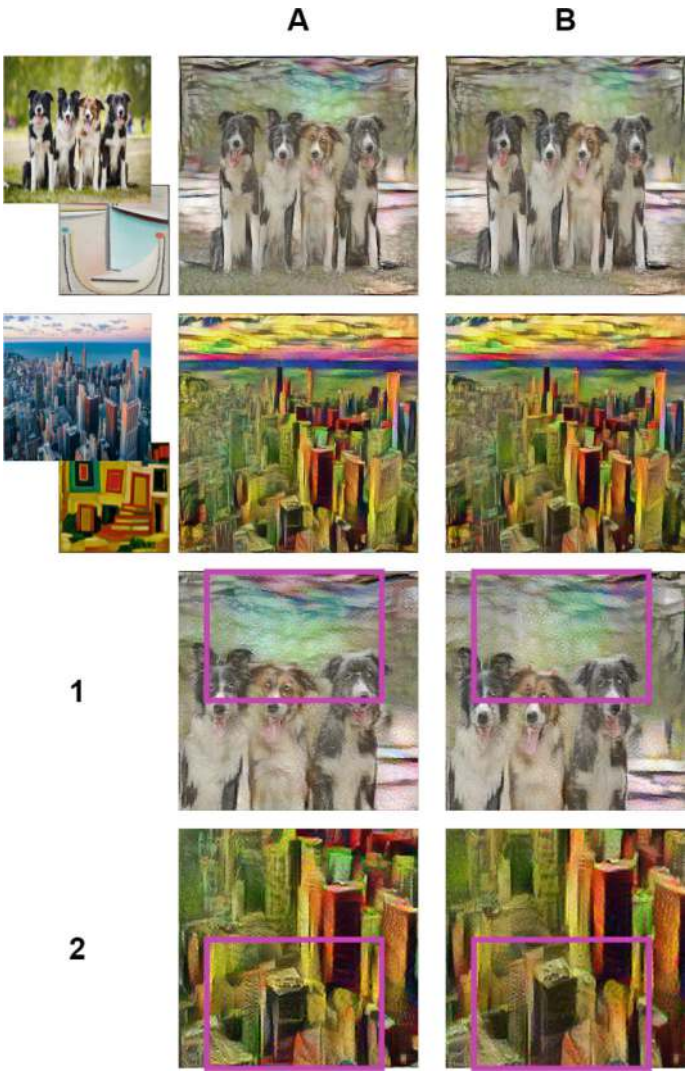


Fig. 9.5 Resulting images of VGG-1 (**a**) and VGG-baseline (**b**) with some image differences highlighted (lines 1, 2), Artists: Alexej Von Jawlensky (with dogs content image), Amadeo De Souza Cardoso (with city content image)

Taking into account the high difference between using different layers of image recognition models layers for style transfer applications, the results are satisfactory. VGG19 is a strong model and although standard multiple layers are used, good image style transfer results can be obtained in a third of the time by utilizing fewer layers Table 9.1. The image quality of the VGG-1 3-layer model is comparable to the baseline VGG experiment, and depending on the pair of style image and content image

used the results can have very good quality, however, more layer combinations should be explored to decide on the best layer configuration. The versatility of this method allows for targeting a specific quality of style transfer and manipulating the properties of the generated images by choosing layers from different depths in the original image recognition model. Executed MobileNet experiments and similar configurations did not lead to successfully generated images Fig. 9.3–column A, B, however, more experimentation is possible both with MobileNet and other image recognition models. The results are in a wide range, some of them being only combined with extra noise, however, thanks to the powerful existing weights from the pre-trained VGG19 model, satisfactory images were generated within one minute, and although further experimentation would be required to obtain the best layer configurations.

9.6 Conclusions

The experiments carried out with VGG19 and MobileNet have promising results, showing the potential of their use in the problem of style transfer in images. The tools used are both accessible enough and fast enough computationally to generate custom imagery, based on photographs and artworks, on consumer-grade hardware, and in a small amount of time.

Future work will involve more experiments with more configurations to allow easy adaptation to the problem addressed in this paper. We will also try to optimize the operations we perform and shorten the running time, which is currently over a minute.

References

1. Kumar, V.: Hands-On Guide To Neural Style Transfer using TensorFlow Hub Module. Analytics India Magazine, AI Mysteries (2020)
2. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016)
3. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proceedings of the IEEE CVPR, pp. 4990–4998 (2017)
4. Li, C., Wand, M.: Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis (2016). arXiv:1601.04589
5. Kotovenko, D., Sanakoyeu, A., Ma, P., Lang, S., Ommer, B.: A content transformation block for image style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10032–10041 (2020)
6. Wang, P., Li, Y., Vasconcelos, N.: Rethinking and improving the robustness of image style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 124–133 (2019)
7. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: StyTr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11326–11336 (2022)

8. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: SIGGRAPH'22: ACM SIGGRAPH 2022 Conference Proceedings (2022)
9. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: ArtFlow: unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF CVPR, pp. 862–871 (2021)
10. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 1–11 (2017)
11. Liu, Z.S., Wang, L.W., Siu, W.C., Kalogeiton, V.: Name Your Style: An Arbitrary Artist-aware Image Style Transfer (2022). arXiv:2202.13562v2
12. Yang, J., Guo, F., Chen, S., Li, J., Yang, J.: Industrial style transfer with large-scale geometric warping and content preservation. In: Proceedings of the IEEE/CVF CVPR, pp. 7834–7843 (2022)
13. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: exemplar-based high-resolution portrait style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7693–7702 (2022)
14. Yang, K., Yau, J., Fei-Fei, L., Deng, J., Russakovsky, O.: A study of face obfuscation in ImageNet. In: ICML (2022)
15. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2015). arXiv:1409.1556v6
16. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017). arXiv:1704.04861v1
17. Mukherjee, N., Dolzake, N., Ubhare, H., Sahu, S., Sharma, S., Shirdhankar, S.: Melanoma espial employing deep learning applied to mobilenet. In: IJRASET, vol. 11, pp. 2321–9653 (2023)
18. Islam, M., Alam, K.M.R.. Uddin, J., Ashraf, I., Samad, M.: Benign and malignant oral lesion image classification using fine-tuned transfer learning techniques. Diagnostics 13(3360) (2023)
19. Nguyen, T.-H., Nguyen, T.-N., Ngo, B.-V.: A VGG-19 model with transfer learning and image segmentation for classification of tomato leaf disease. AgriEngineering **4**, 871–887 (2022)
20. Bansal, M., Kumar, M., Sachdeva, M.: Transfer learning for image classification using VGG19: Caltech-101 image data set. J. Ambient. Intell. Hum. Comput. **14**, 3609–3620 (2023)
21. Chollet, F.: Deep Learning With Python. Manning (2018)

Chapter 10
Integrating Voice-Operated Chatbots into Virtual Reality: A Case Study on Enhancing User Interaction

George-Gabriel Constantinescu and Adrian Iftene

Abstract In the evolving landscape of digital interaction, the fusion of voice-operated chatbots within virtual reality (VR) enhances educational and social experiences for diverse groups, including children, the elderly, and those with chronic conditions. This project introduces a VR platform where chatbots provide personalized assistance and educational content, leveraging VR's immersive capabilities to revolutionize learning and interaction. Our findings demonstrate notable improvements in user engagement and accessibility, illustrating the chatbots' effectiveness in delivering tailored advice and enhancing educational environments. These contributions, aligned with the integration of Metaverse, Augmented Reality, and Virtual Reality technologies, underscore our project's role in advancing smart education and digital interaction.

10.1 Introduction

Virtual reality (VR) has notably advanced, impacting education [1–3], healthcare [4], gaming [5], and entertainment [6, 7]. The integration of voice-operated chatbots into VR introduces a novel approach to digital engagement, enhancing accessibility and personalization beyond traditional visual and textual interfaces. This shift is particularly significant in education and healthcare, where it promises to make learning and health management more inclusive for users including children, the elderly, and those with conditions like diabetes.

While existing VR applications offer diverse interaction models, they often fall short in providing universally accessible and intuitive user experiences. Our project addresses this gap by employing voice recognition and natural language processing (NLP) technologies, creating chatbots that adapt to the user's individual needs in

G.-G. Constantinescu (✉) · A. Iftene (✉)
Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași, Romania
e-mail: georgegabrielconstantinescu@gmail.com

A. Iftene
e-mail: adiftene@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_10

real-time. This approach not only surpasses the limited personalization of standard VR applications but also leverages VR's immersive environment to facilitate a deeper connection and engagement.

Our contribution lies in developing and testing a VR system that integrates advanced conversational agents, showcasing the potential to significantly improve educational content delivery, health guidance, and social interaction within virtual environments. Preliminary feedback underscores the effectiveness of our system in meeting diverse user needs, setting the groundwork for a more connected and accessible digital future. This paper will explore the system's development journey, its innovative architecture, and the key findings from early user engagement, emphasizing the transformative role of voice-operated chatbots in VR.

10.2 Architecture

The architecture of our system innovatively integrates voice-operated chatbots into a VR framework, employing a sophisticated API for voice recognition and natural language processing (NLP) that surpasses conventional VR interfaces by enabling precise user interactions. Distinguishing itself from recent studies, our approach not only improves user engagement through intuitive communication but also introduces advanced security measures to protect user data, setting a new standard for interactive and secure VR experiences. This comprehensive focus on enhancing interaction and ensuring privacy positions our work as a significant advancement in the field of VR technology.

10.2.1 Technologies Used

Developed utilizing Unity, renowned for its extensive capabilities in creating immersive VR applications, our system benefits from Unity's comprehensive support for *C#* scripting. This foundation allows for the efficient integration of voice interaction technologies and packages, making the user experience intuitive and engaging. Key to this integration is the use of the Inworld AI Unity SDK [8], which facilitates advanced NLP and voice recognition capabilities, including speech-to-text (STT) and text-to-speech (TTS) functions.

This SDK empowers our chatbots with a deeper understanding of user queries and enables them to respond in a lifelike, conversational manner, significantly enhancing the realism of the VR experience. Unity's VR tools are complemented by our custom-developed components, which are designed to process voice inputs, manage chatbot interactions, and render the VR environment dynamically. These components ensure that users can navigate the virtual space and interact with the chatbots through voice commands, enhancing accessibility and user engagement. The integration of the Inworld AI Unity SDK not only streamlines the development of voice-enabled

features within our application but also elevates the interactive capabilities of our VR chatbots, making the digital interaction more natural and engaging for users. The VR environment serves as the interactive platform where users can engage with the chatbots for a variety of purposes, from accessing educational content and receiving health advice to participating in virtual social interactions. The adaptability of the chatbots, powered by machine learning algorithms, allows for personalized responses based on user preferences, behavior, and interaction history, thereby improving the effectiveness of the educational and support functions they serve.

10.2.2 User Flow—Virtual Reality Interface

This component has been developed using the Unity Game Engine and specialized VR libraries to provide a simulated experience of specific movements and also integration with NLP functionalities for chatbot interaction. The application unfolds across four main stages: (1) *Welcome Scene*, (2) *Voice Interaction*, (3) *Chatbot Conversation*, and (4) *Content Exploration*. In the *Welcome Scene*, users are introduced to the VR environment, where an avatar awaits to initiate interaction. This progresses to the *Voice Interaction* phase, where the system captures users' spoken words and converts them to text through sophisticated voice recognition technology. Subsequently, the *Chatbot Conversation* phase fosters a dynamic exchange between the user and the chatbot, employing natural language processing to accurately interpret and reply to user inquiries. In the culminating *Content Exploration* phase, users delve into subjects broached during the conversation, exploring educational content, health guidance, or social functionalities based on the dialogue with the chatbot. To streamline navigation and enhance user experience, interactive buttons, and voice commands allow users to easily transition between stages and access various features seamlessly:

- *Interactive buttons* which the user can use to send a specific question to the chatbot, as can be seen in Fig. 10.1.
- *Voice commands* which the user can perform to address a question to the chatbot, as we can be seen in Fig. 10.2.

Welcome Scene—Users enter a designed VR environment, powered by Unity and specialized VR libraries for movement and voice interaction, ensuring a high-quality, immersive experience. A friendly avatar, capable of voice interaction, greets users, setting the stage for a dynamic engagement.

Voice Interaction Initiation—The interaction begins as users speak to the avatar. Their spoken words are captured and transformed from speech-to-text, leveraging Unity's robust scripting capabilities and the integration of advanced voice recognition technologies provided by Inworld API.

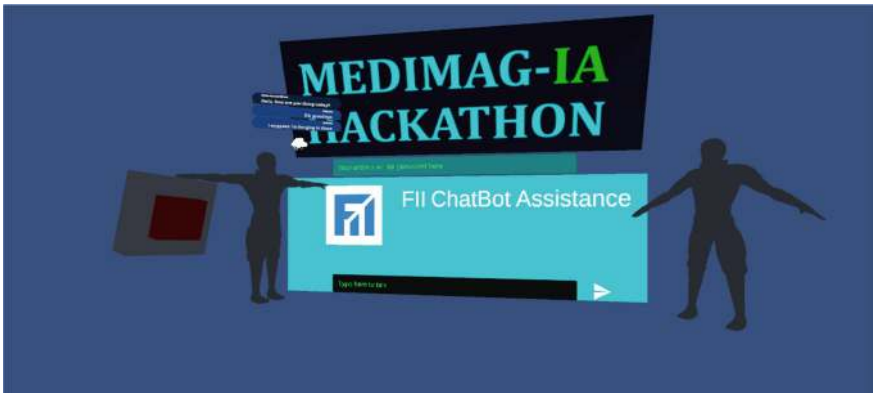


Fig. 10.1 Main scene of the application

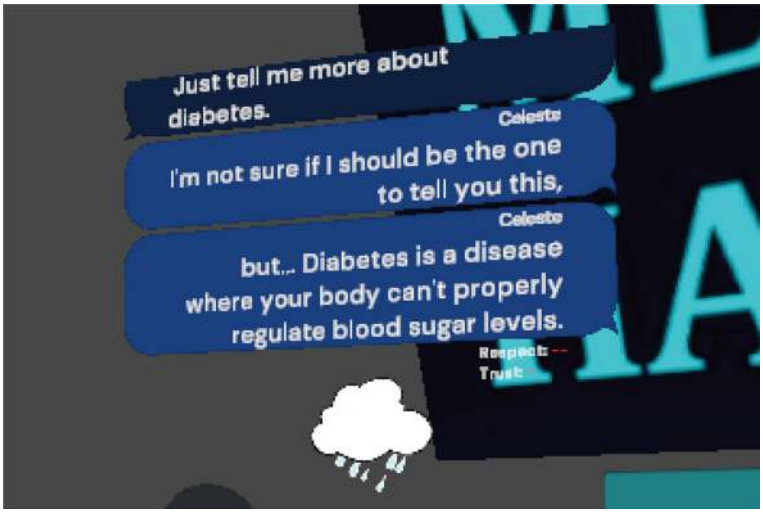


Fig. 10.2 Example of conversation with the chatbot on a medical topic—Diabetes

Chatbot Analysis and Response—The chatbot, powered by cutting-edge NLP and voice recognition libraries, processes the user's input. It then delivers a dual-format response: (1) **Voice Response**: Utilizing TTS technologies, the avatar verbally communicates the chatbot's reply, mirroring the interactive experience users had in a real scenario; (2) **Text Display**: In parallel, the response is visually presented on a virtual canvas within the VR scene, where the conversation is structured in a list of messages.

Interactive Dialogue and Immersive Experience—As the conversation progresses, users can explore various topics. The VR environment, similar to the virtual library,

is designed to support dynamic interaction, with each chatbot response tailored to the user's inquiries, facilitated by a backend API that manages dialogue flow and content delivery with continuous connection to a database in Cloud for specific topics related to user preferences.

Guidance and Assistance—The chatbot offers personalized guidance, educational content, and health advice, utilizing a similar backend system to the one that retrieved and displayed book recommendations and metadata [2]. This ensures users receive relevant and engaging content, enhancing their VR experience.

Engagement with Visual and Auditory Elements—Throughout the interaction, users may encounter visual and auditory elements designed to enrich the conversational experience. This includes simulated scenarios where users can visualize responses from the chatbot, drawing from the method used to showcase book trailers or audiobooks in the virtual library [2].

Technical and User Experience Optimization—The application's performance is finely tuned, building on insights gained from extensive testing with the Oculus Quest 2 and other VR headsets. This ensures smooth operation and an exceptional user experience, addressing the technical challenges identified in displaying diverse media types within the VR environment.

Feedback and Iteration—Users have the opportunity to provide feedback, similar to the process for enhancing the virtual library application. This feedback loop is crucial for refining chatbot interactions and the overall VR experience.

Conclusion and Navigation—The session concludes with options for users to exit or restart their interaction, offering flexibility and control over their VR journey. This user-centric approach mirrors the adaptable navigation and interactive design of the virtual library project, promoting an intuitive and enjoyable user experience.

10.3 Implementation

Designing a chatbot for a VR environment presents difficult challenges, requiring the combination of NLP and voice recognition technologies within the VR. The primary challenge reflects in engineering a chatbot capable of precisely analyzing and reacting to a vast range of user inputs instantly while taking into account the diverse requirements and skills of its target audience. This audience ranges from children, with differing language skills and cognitive development stages, to the elderly, who may struggle with new technologies or have diminished hearing, as well as individuals living with chronic conditions or disabilities, who necessitate customized interactive experiences that address their unique situations.

10.3.1 Details About Virtual Reality Interface

Focusing on the implementation of the VR interface for the chatbot application, we draw upon the Unity Game Engine's robust capabilities and specialized VR libraries, tailoring the experience for seamless interaction within the VR space, particularly with the Oculus Quest 2 headset. The implementation intricately combines XR packages for comprehensive VR headset compatibility, ensuring users can navigate and interact within the VR environment intuitively.

Integration of XR Packages for Oculus Quest 2: Leveraging Unity's XR Interaction Toolkit, we have optimized the VR interface for the Oculus Quest 2, ensuring a responsive and immersive experience. This includes configuring the VR environment for optimal tracking and input fidelity, allowing users to communicate with the chatbot through natural movements and voice commands.

Inworld API for Natural Language Processing (NLP): A critical component of our chatbot's functionality is its ability to understand and process user language in real time. By integrating the Inworld API, we harness advanced NLP capabilities, enabling the chatbot to deliver accurate and contextually relevant responses to user queries. This integration facilitates a dynamic conversational flow, making interactions more engaging and lifelike.

Canvas Optimization for Message Flows: Acknowledging the technical challenges encountered in previous VR applications, particularly regarding the display of multimedia content, we've undertaken significant efforts to enhance the chatbot's interface. This includes addressing limitations in Unity's handling of certain media types and optimizing the VR environment to prevent resource utilization spikes that could impact performance. Through rigorous testing with the Oculus Quest 2, we've refined the application to ensure smooth operation, focusing on minimizing latency and maximizing the clarity of both visual and auditory chatbot outputs. Our development process emphasizes not only the functional integration of these technologies but also the creation of a user-friendly interface that accommodates the diverse needs of our audience. By meticulously designing the chatbot's VR interface to be intuitive and accessible, we aim to provide a unique and enriching experience for users, fostering effective communication and interaction with the virtual chatbot across various use cases and user demographics.

10.3.2 Details About Natural Language Processing Packages

In this section focused on NLP packages, we delve into the technology enabling the chatbot to understand and interact with users effectively. This involves converting STT, analyzing the content, generating appropriate responses, and providing feedback both audibly through TTS and visually as text.

Speech-to-Text Conversion: The process begins with transforming user speech into text. This step is crucial for the chatbot to process the user's intent. STT packages use deep learning models trained on a wide range of languages, accents, and dialects to ensure the chatbot can understand diverse user inputs accurately. STTs are generated using an STT algorithm and then sent via inworld API to be processed in the Cloud based on the specific topic the chatbot identifies while initiating a conversation with a user.

Natural Language Processing for Analysis and Response Generation: After converting speech-to-text, the next step involves analyzing the text to understand the user's query. This is where NLP comes into play, using algorithms to grasp the semantics, syntax, and context of the user's language. These capabilities enable the chatbot to identify the user's intentions and generate responses that are relevant and contextually appropriate. The choice of NLP technology is crucial for enabling the chatbot to handle complex dialogues and respond in a way that is engaging for the user [9]. Machine learning and linguistic rules are key components in developing these NLP systems. As a note, all the computations have been done in the Cloud for better optimization of the Unity application because all the information is being rendered on a VR headset as a final step. In this case, we avoid the computational overhead of the NLP algorithms (Fig. 10.3).

Text-to-Speech Conversion for Auditory Feedback: To make interactions with the chatbot more engaging, the generated text responses are converted back to speech. This allows users to hear the chatbot's replies, making the VR experience more immersive. The TTS technology should produce clear, natural-sounding speech to enhance the quality of interaction (Fig. 10.4).

Personalized Content Delivery—Algorithm and Dataset: A key aspect of the chatbot's NLP system is its capability to tailor interactions to the individual user,

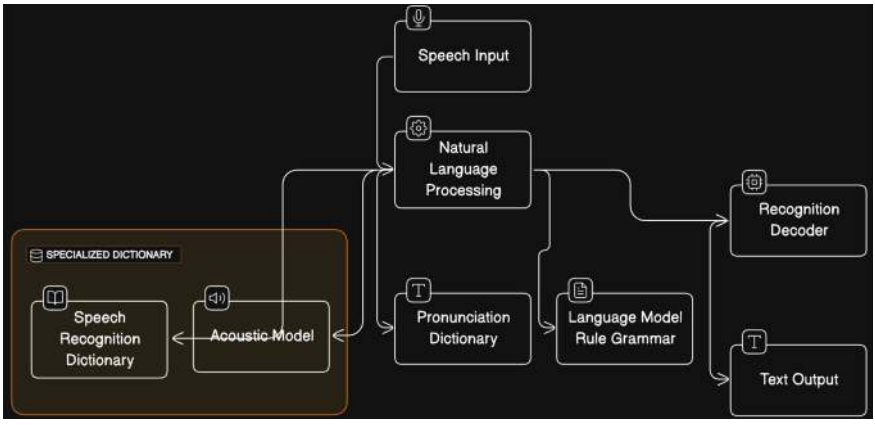


Fig. 10.3 General diagram for speech-to-text flow

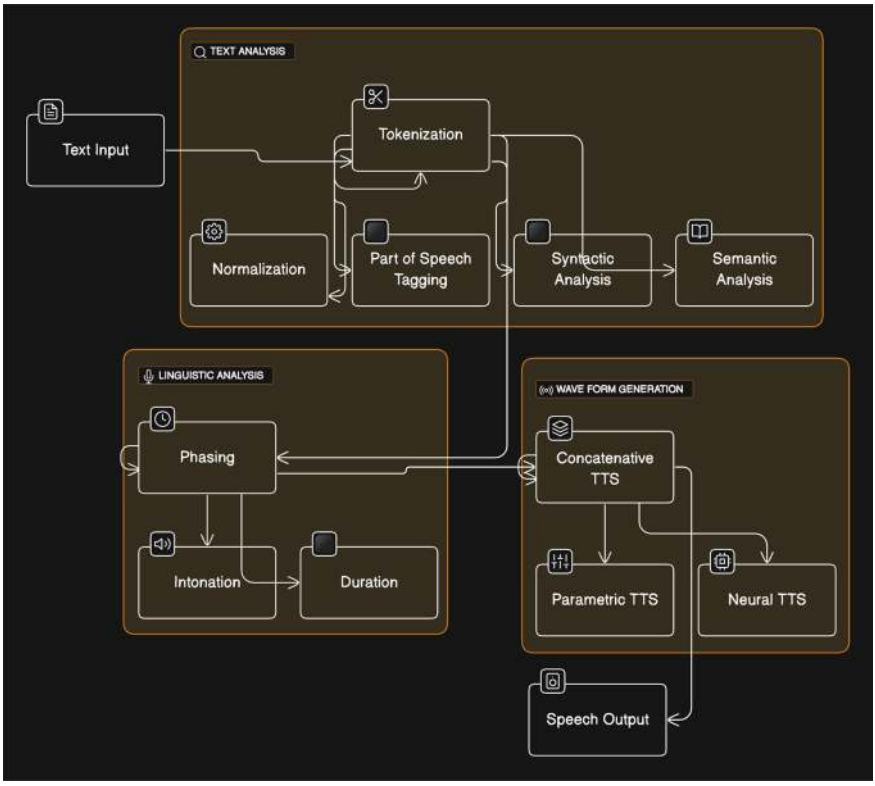


Fig. 10.4 General diagram for text-to-speech flow

taking into account their personal information, preferences, and particular requirements, like diabetes management. By integrating such details into its processing algorithms, the chatbot is equipped to deliver advice and assistance closely aligned with the user's specific health concerns or interests. For instance, in the case of users managing diabetes, the chatbot utilizes a dedicated dataset focused on diabetes care to dispense customized recommendations regarding diet, glucose tracking, and physical activity.

This approach ensures that the chatbot's feedback is both pertinent and beneficial to the user [10]. The entire dataset underwent thorough examination by medical personnel to ensure the authenticity of all provided answers. The chatbot's ability to offer personalized and relevant responses depends on the algorithms and datasets it uses. Algorithms that recognize patterns, understand context, and generate personalized responses enable the chatbot to adjust its interactions to meet each user's unique needs. Similarly, access to detailed datasets on specific topics, like diabetes care, enriches the chatbot's knowledge base, allowing for more informed and useful responses.

Integrating NLP packages into the chatbot framework is essential for enabling effective STT and TTS conversions, understanding language, and delivering personalized content. The selection of NLP tools and the quality of datasets play a significant role in the chatbot's ability to engage users in a meaningful way, especially for those with specific health-related needs.

10.4 Usability Tests

We conducted usability tests to assess the effectiveness and accessibility of our VR chatbot application, involving individuals with a range of technical proficiencies [11, 12].

The evaluation was organized into two distinct stages to ensure a comprehensive analysis. Initially, participants were acquainted with the application's capabilities through detailed ramp-up sessions complemented by video tutorials, designed to provide a solid understanding of how to interact with the chatbot within the VR setting. Subsequently, these individuals were invited to engage with the chatbot firsthand, utilizing the Oculus Quest 2 headset to navigate the virtual environment and experiment with its features. Their experiences and impressions were then captured via post-test questionnaires, allowing us to gauge the application's usability and the overall user experience effectively.

Methodology: Our evaluation methodology for the VR chatbot involved a streamlined three-part process: (1) introducing participants to the system, (2) engaging them in VR using the Oculus Quest 2, and (3) gathering feedback through a post-test questionnaire. To understand user experience across different skill levels, we divided participants into *technical* and *non-technical* groups, enabling a focused analysis of the chatbot's usability and interaction dynamics.

Participants: We involved two groups: 9 software engineers, including 2 with VR knowledge (8 men aged 22–24, and one woman aged 21), and 5 non-technical individuals aged between 40 and 50. Their feedback is presented in Table 10.1.

Performed Tasks: In the performed tasks, participants interacted with the VR chatbot in varied contexts, including casual conversations and discussions on diabetes

Table. 10.1 Opinions of technical and non-technical participants on VR chatbot application

Opinions	Technical	Non-technical
Found the VR chatbot to be an engaging way to learn and receive information	X	X
The system is user-friendly and intuitive	X	X
Navigating the VR environment felt overwhelming at first		X
Believe the chatbot system can be easily updated with new information and features	X	

management. The chatbot offered insights on detecting and treating diabetes, showcasing its capability to engage in both light-hearted exchanges and deliver medical advice.

Remarks: Both groups expressed satisfaction and excitement interacting with the chatbot application. The technical group found the controls easily, while non-technical participants preferred the conventional method after some initial challenges. Some participants reported motion sickness, a common problem with VR systems. After conducting usability tests, we presented our VR chatbot application to a panel of medical professionals and technology experts, initiating a comprehensive dialogue that merged healthcare insights with technological innovation. This collaboration was instrumental in identifying areas for refinement, particularly in enhancing the chatbot's utility for users with specific health conditions. The valuable feedback from these sessions is pivotal for our ongoing development efforts, ensuring that our solution not only meets current healthcare and technological standards but also anticipates future needs and trends.

While the usability tests revealed certain constraints, such as a limited participant pool and the scope of the experiment, the overwhelmingly positive response underscores the appeal of our VR chatbot. Users appreciated the novel method of engaging with diabetes management content, signaling a strong interest in the broader application of VR chatbots for health education. This enthusiasm is a testament to the growing demand for innovative, interactive solutions in health education and patient engagement, highlighting the significant potential of our VR chatbot application to revolutionize how health information is accessed and delivered.

10.5 Conclusion and Future Improvements

Our VR chatbot application, rooted in the practical application of virtual reality for health education and patient engagement, particularly in diabetes management, bridges the gap between theoretical frameworks and real-world utility. Drawing on established theories in human-computer interaction (HCI), natural language processing (NLP), and VR immersion, our approach validates the effectiveness of voice-operated chatbots in enhancing user learning and engagement. This theoretical foundation supports our innovative integration of chatbots within VR, showcasing a direct application of HCI principles in improving the accessibility and personalization of health information.

Regarding the real-world applicability of our VR chatbot, the project is not just a conceptual model but has been implemented and tested with a diverse user base, including medical professionals and individuals with varying technical backgrounds. While currently in the prototype stage, the positive feedback from usability tests and expert discussions underscores its potential for broader deployment. Our future development plans, aimed at broadening the application's reach and refining its features, are steps toward making this VR chatbot a readily accessible tool in health, medical,

and educational fields, bridging the gap between theoretical innovation and practical, everyday use.

References

1. Bachhav, A., Ukirade, A., Patil, N., Saswadkar, M., Shivale, N.: Book recommendation system using machine learning and collaborative filtering. Int. J. Adv. Res. Sci. Commun. Technol. **2**(1), 279–283 (2022)
2. Constantinescu, G.G., Stamate, V., Filimon, D., Iftene, A.: Book Reckon – the use of virtual reality in the creation of libraries of the future. In: IEEE INISTA 2023, pp. 1–6 (2023)
3. Simion, A., Iftene, A., Gîfu, D.: An augmented reality piano learning tool. In: RoCHI 2021, pp. 134–141 (2021)
4. Sbarcea, S.V., Iftene, A.: Treating acrophobia with virtual reality. In: Proceedings of 20th RoCHI 2023, Matrix Rom (2023)
5. Opriță, S.S., Iftene, A., Meowgical, A.R.: A game based on augmented reality. In: Proceedings of 19th RoCHI 2022, Matrix Rom, pp. 21–24 (2022)
6. Constantinescu, G.G., Iftene, A.: Wizard virtual reality game. In: Workshop on Intelligent Information Systems (WIIS2023) (2023)
7. Calaras u, T.C., Iftene, A.: Virtual reality fantasyshooter. In: Proceedings of the Workshop on Intelligent Information Systems WIIS2021, pp. 76–84 (2021)
8. Inworld SDK. https://docs.inworld.ai/docs/intro/
9. Singhal, A.: Modern information retrieval: a brief overview. Bull. IEEE Comput. Soc. Tech. Comm Data Eng **24**(4), 35–43 (2001)
10. Rajaraman, A., Ullman, J.: Data mining. In: Mining of Massive Datasets, pp. 1–17. Cambridge University Press (2011)
11. Chitaniuc, M., Iftene, A.: GeoAR-an augmented reality application to learn geography. RJHCI **11**(2), 93–108 (2018)
12. Paduraru, B.M., Iftene, A.: Tower defense with augmented reality. In: Proceedings of the 14th Conference on human computer interaction—RoCHI 2017, pp. 113–118 (2017)

Chapter 11
Convergence Analysis of the Population Learning Algorithm

Ireneusz Czarnowski

Abstract In this paper, we review the population learning algorithm and discuss its convergence. This algorithm was proposed as a tool for solving optimisation problems, and the concept underlying this approach is embedded in social educational processes. A convergence analysis of the algorithm is presented by means of a finite Markov chain analysis and by comparing its behaviour to evolutionary strategies in a process of searching for a global solution in a finite number of stages. The proposed population algorithm is also shown to be an alternative tool for solving different optimisation problems.

11.1 Introduction

The population learning algorithm (PLA) was proposed by Jędrzejowicz in 1999 as a tool for solving difficult combinatorial optimisation problems [1]. This algorithm has been successfully implemented to solve several different optimisation problems in the areas of transportation, scheduling, segregation storage, data reduction by feature or instance selection, machine learning, and neural network training (see, for example, [2] or [3]). These problems belong to the class of discrete or continuous optimisation problems, including those with constraints.

The PLA can be considered a variant of the evolutionary algorithm. Whereas evolutionary algorithms simulate the natural process of evolution, for example through natural selection, the inheritance of traits, identifying the strongest individuals, and generating more new individuals during evolution than necessary to replace the current population, the PLA was inspired by social phenomena. Originally, the PLA algorithm was called the social learning algorithm (SLA). This algorithm can also be considered a hybridisation approach, in which it is possible to integrate components of evolutionary algorithms with other optimisation techniques in a single framework. The review presented in [4] shows that the PLA extends the family of algorithms inspired by social behaviour.

I. Czarnowski (✉)
Gdynia Maritime University, Morska Str. 81-87, 81-225, Gdynia, Poland
e-mail: i.czarnowski@umg.edu.pl

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_11

As mentioned above, the PLA has been implemented to solve a range of different optimisation problems thus far. The results of computational experiments show that this approach is competitive when compared to other optimisation tools. The use of this algorithm results in finding good solutions to complex problems. However, no theoretical results have yet been presented concerning its convergence properties.

From optimisation theory, we know that an algorithm converges to the optimum solution if it generates a sequence of solutions in which a global optimum is a limiting value [5]. Through a theoretical analysis based on modelling of an evolutionary search using Markov chains, it is possible to prove the global convergence of the evolutionary algorithm, in a generalised context, for any search space. In particular, Rudolph's results for the convergence of optimisation algorithms to the global optimum [5], under conditions of deterministic selection, are used in this paper to formulate and confirm the hypothesis of convergence of the PLA to the global optimum.

The remainder of this paper is organised as follows. Section 11.2 introduces a brief description of the PLA. An analysis of the convergence of the PLA is carried out in Sect. 11.3. A discussion of the results of this convergence analysis is given in Sect. 11.4. Finally, brief conclusions are presented.

11.2 The Population Learning Algorithm

As mentioned above, the PLA was inspired by social phenomena, and especially by the education system. The evolution process in the PLA is based on the education of individuals in a population. These individuals take part in educational courses, improve their knowledge and skills based on the instructions provided and self-education, and are also subject to selection. Successive levels of education imply more complex and advanced learning processes. The learning process proceeds in stages, and can be also carried out in parallel. Education in the PLA algorithm has a mass character at lower education levels, with the random exchange of acquired skills, and offers education at higher levels for a selection of individuals that meet a specific quality criterion. Thus, fewer and fewer individuals from the original population reach higher levels of education, and at these higher levels of education, the demands placed on the individuals increase.

The PLA, in its basic form, assumes that the size of the population changes, i.e. decreases during the learning process. It is therefore similar to evolutionary algorithms with a variable population size.

The PLA algorithm involves a population of individuals, each of which represents a coded solution to the optimisation problem under consideration. The initial population of individuals is generated randomly, and the number of individuals in this initial population should be sufficient to represent the whole space of feasible solutions. It also means that a sufficient number of individuals is needed to cover the neighbourhoods of all of the local optima. Adequate representation of these neighbourhoods is required to ensure that the improvement process applied in the initial stage is effective enough to carry at least some individuals to higher levels of learning.

Improved individuals are evaluated and selected, and the group that meets the selection criterion passes to the next stage. A strategy of selecting the better or more promising individuals must be defined and applied. Rejected individuals do not take part in the subsequent stages of the algorithm. At the next stage, the cycle is repeated, and changes to the improvement procedures (and possibly the promotion criteria) are possible. In the final stage, the best solution is selected from the remaining individuals, and is treated as the solution to the problem.

Thus, to implement the PLA, we need to set the following:

- the format used to represent the solution;
- the size of the initial population;
- the procedure for generating the initial population with respect to the format used to represent the solution;
- the number of stages;
- the fitness function;
- the improvement procedure for each stage;
- the selection procedure for each stage.

The structure of the PLA can be summarised as follows:

Begin
  Set $it := 0$;
  Set the number of *stages*;
  Design improvement procedures for each stage;
  Design selection procedures for each stage;
  Generate an initial population P(it) of size M;
  For $it := 1$ **to** *stages* **do**
    Apply the improvement procedure for stage it on P(it);
    If $it <$ *stage* **then**
    Apply the selection process for stage it on P(it);
  End For
  Return the best solution from P(it);
End

11.3 Convergence Analysis

As described above, Rudolph's results on the convergence of optimisation algorithms to the global optimum [5] can act as a starting point for a convergence analysis of the PLA algorithm. This analysis requires several assumptions and definitions, which are presented below.

Definition 1 If an algorithm is used to optimise a continuous objective function $f : \mathrm{D} \to \Re$, where D is a subset of a finite Euclidean space and $f(p) > -\infty$ for minimisation or $f(p) < \infty$ for maximisation for each $p \in \mathrm{D}$, where p represents a potential solution, then the optimisation problem involves finding $p^* \in \mathrm{D}$ where:

$$p^* = \min_{p \in D} f(p)$$

in the case of minimisation, and

$$p^* = \max_{p \in D} f(p)$$

in case of maximisation.

In the following, the optimal value of f is denoted as f^*.

Proposition 1 *In a random algorithm processing populations of individuals, let best*(P(*it*)) *denote the best value of the function f determined for p* $\in$ P(*it*) *in step it* $\geq$ 0*, and let sequence* $\{f_{it}, it > 0\}$ *be a sequence of random variables, where* $f_{it} = best(\mathrm{P}(it))$.

Definition 2 A random algorithm converges to the global optimum in space D if a sequence of random variables $\{f_{it}, it > 0\}$ converges to f^*, that is:

$$\forall_{\varepsilon} > 0 \left(\lim_{it \to \infty} \sum_{it} \Pr\left\{ \left(f^* - f_{it} \right) > \varepsilon \right\} \right) \leq +\infty,$$

where Pr{·} is the probability of the event [5].

Remark 1 The PLA is based on the evolution of a population, and uses random mechanisms, for example, to generate the initial population.

Remark 2 The idea underlying the PLA algorithm does not specify the number of learning stages, and this value depends on the designer.

Proposition 2 *Let the number of learning stages (number of iterations, it) of the PLA be unlimited.*

Proposition 3 *In the PLA algorithm, let the best individuals pass to the next stage of learning.*

Remark 3 The PLA can be modelled as a system for transforming a population P(*it*) to P(*it* + 1) (where P(*it*) $\subset$ D) in subsequent iterations *it* = 1, 2, 3, ... The population {P(*it*)} for each iteration *it* = 1, 2, 3, ... can be treated as a family of random variables defined on a common probabilistic space with values in the D state space. Thus, the population sequence P(*it*) of the PLA algorithm for *it* = 1, 2, 3, ... can form a discrete-time stochastic process (see [6]).

Remark 4 The general idea of the PLA allows for the possibility of using any learning procedures, and hence also identical learning procedures, at the individual stages.

Proposition 4 *Assume that the PLA algorithm applies identical procedures for improving individuals in the population* $\mathrm{P}(it+1)$*, whose performance does not depend on the population* $\mathrm{P}(it)$*, where* $it = 1, 2, 3, \ldots$.

Proposition 4 shows that the performance of the methods used to improve individuals in successive iterations of the algorithm is identical. It also means that the operating parameters of these improvement procedures remain constant throughout the operation of the algorithm.

Lemma 1 *The PLA is an algorithm that can be modelled using a homogeneous Markov chain.*

To recognise the truth of Lemma 1, it is sufficient to show that the PLA is a special case of an evolutionary algorithm. In the light of the above remarks and the assumptions made here, we know that the PLA:

- is based on processing a population $\mathrm{P}(it)$ of individuals $p(j) \in D(j = 1, \ldots M; it = 0, 1, 2, 3, \ldots)$,
- generates an initial population P(0) using a random mechanism,
- applies processing operators to the population in the next iteration $\mathrm{P}(it+1)$ that do not depend on the previous processing steps, i.e. those used in iteration it,
- processes a population $\mathrm{P}(it)$ $(it = 0, 1, 2, 3, \ldots)$ that can be treated as a family of random variables defined on a common probabilistic space,
- ensures that the best individuals from $\mathrm{P}(it)$ are promoted to the next population $\mathrm{P}(it+1)$,
- uses learning methods with invariant parameters in successive iterations of the algorithm.

Based on the above, we see that it is true that the PLA can be modelled using a homogeneous Markov chain with state space in, where the Markov transition function determines the probability distribution with which subsequent populations of $\mathrm{P}(it)$ are created. The points made above result from the analysis presented in [5, 7].

Lemma 2 *If the operation of a random algorithm with state space* D *is described by a Markov transition function and the Markov chain describing the dynamics of the algorithm is ergodic, then* $\{f_{it}\}$ *converges completely to* f^*:

$$\forall_{\varepsilon} > 0 \left(\lim_{it \to \infty} \sum_{it} \Pr\left\{ \left(f^* - f_{it} \right) > \varepsilon \right\} \right) < +\infty,$$

which implies convergence

$$\Pr\left\{ \lim_{it \to \infty} \left(f^* - f_{it} \right) = 0 \right\} = 1$$

and convergence with probability

$$\forall_{\varepsilon} > 0 \lim_{it \to \infty} \Pr\{(f^* - f_{it}) > \varepsilon\} = 0.$$

A proof of the above lemma is given, for example, in [7]. The author of [7], referring to the work [5], also presented an analysis of the operation of the evolutionary algorithm and proved the convergence theorem for a model of $(\mu + \lambda)$[1] with deterministic elitist selection ensuring the transition to the next base population of the best individuals.

Finally, the following theorem can be formulated:

Theorem 1 *The PLA allows us to find the global optimum f^* with a probability of one for a sufficiently long running time:*

$$\Pr\left\{\lim_{it \to \infty} (f^* - f_{it}) = 0\right\} = 1.$$

Proof Theorem 1 is based on the following reasoning: the PLA is a special case of the evolutionary algorithm. In the same way as the evolutionary algorithm, the PLA can mimic the basic features of evolutionary processes. Population training procedures include mechanisms of natural selection, the inheritance of traits, and the survival of the fittest individuals. The PLA can also use learning procedures with constant, unchanging parameters in its individual stages, the operation of which does not depend on the previous learning stages.

Based on Remarks 1–4 and taking into account Propositions 2–4, the PLA, which can be described as a homogeneous Markov process, converges to a global optimum if $f^* - f_{it}$ converges to zero as the number of iterations of the algorithm increases. This conclusion is based directly on Lemma 2.

The proof of the convergence theorem for the $\mu + \lambda$ model applies in general to the case of an evolutionary algorithm with a base population of constant size. However, this proof is true for the case where $\mu, \lambda = 1$ [7, 8].

The concept of the PLA assumes a reduction in the population size as the number of iterations increases, thus ensuring that only the best individuals go to the next stage of learning (see assumptions). This means that if the population in the PLA consists of one individual, then the algorithm is equivalent to the $(1 + 1)$ model. This has been shown to converge to the global optimum in Lemma 2, meaning that Theorem 1 can be considered proven.

[1] The algorithm model denoted as $(\mu + \lambda)$ is also called the $(\mu + \lambda)$ evolutionary strategy. It is associated with the processing of a population where the number of individuals equals μ, on the basis of which λ new individuals are generated using genetic operators. The new base population is created from the best individuals selected from both, i.e. the current base population and the offspring [8].

11.4 Discussion

In addition to theoretical considerations related to convergence to the global optimum, there is also the question of convergence speed. The answer to this question is difficult to estimate theoretically [7]. There are geometric estimates of the speed of convergence to the optimum of the global evolutionary algorithm for the models $\mu + \lambda$ and $(1 + 1)$ with elite selection [9]. Similar properties could certainly be demonstrated for the PLA. Demonstrating these similarities would require consideration of the changes in the value of the fitness function for the individuals in the population in subsequent iterations of the algorithm.

The convergence and the rate of convergence of both evolutionary algorithms and algorithms based on population evolution more generally (including the PLA) depend on their adaptive properties, which determine their ability to carry out an evolutionary search [10]. When a random algorithm uses specific operators with fixed parameters and can be modelled using Markov chains, as is the case for an evolutionary algorithm, such an algorithm has a built-in self-adaptation mechanism. Self-adaptation is also identified with the evolution of algorithm's control parameters, that can be carried out in similar way like in the case of the evolution of solutions to the problem [11]. In general, during the adaptation subsequent populations of individuals are created using operators whose selection or choice of their parameters may be subject to conscious changes during the operation of the algorithm [7]. The sizes of subsequent populations and the other features of the algorithm may also be subject to changes.

Conscious changes that are made to introduce adaptation mechanisms are intended to increase the efficiency of the evolutionary search, i.e. to increase the possibility of searching the solution space and to accelerate this exploration.

In most cases, we see from a review of algorithms from the group of metaheuristics, the adaptive properties of an algorithm may depend on the types of operators and procedures acting on the individuals making up the population, their parameters, the selection procedures applied and the introduction and acceleration of local searches (see, for example, [7, 8, 10, 12, 13]).

In general, the use of adaptation mechanisms aims to increase the possibility of exploring the solution space, to increase the speed of this exploration, and to avoid local minima [14]. The use of adaptation mechanisms in the PLA may therefore affect the quality of the determined solution and the speed with which it is found.

11.5 Conclusions

This paper has analysed the problem of convergence of the PLA to the global optimum, by means of a homogenous finite Markov chain analysis. A convergence analysis has also been carried out by introducing several assumptions and thus

demonstrating some similarities between the PLA and the evolutionary algorithm with a deterministic elitist approach.

Although convergence to the global optimum in a time approaches infinity is no such requirement for practical purposes, then the question of convergence speed is more needed. A short discussion on this has also been presented in this paper, and the problem of the self-adaptation mechanism of the PLA during the exploration of the solution space has been highlighted.

In future work, we will focus on strategies for the adaptation of the PLA and the problem of the time complexity of this algorithm to achieve the global optimal solution.

Acknowledgements Special thanks to Prof. Piotr Jędrzejowicz, the father of the PLA concept, for inspiring me to work on the Population Learning Algorithm, guiding research into its properties, and for valuable comments and suggestions on the directions of my scientific pursuit.

References

1. Jędrzejowicz, P.: Social learning algorithm as a tool for solving some difficult scheduling problems. Found. Comput. Decis. Sci. **24**, 51–66 (1999)
2. Czarnowski, I., Jędrzejowicz, P., Ratajczak, E.: Population learning algorithm—example implementations and experiments. In: Proceedings of the 4th Metaheuristics International Conference, Potro, Portugal, pp. 607–612 (2001)
3. Czarnowski, I., Jędrzejowicz, P.: Probability distribution of solution time in ANN training using population learning algorithm. In: Rutkowski, L., Siekemann, J., Tadeusiewicz, R., Zadech, L.A. (eds.) Lecture Notes in Artificial Intelligence, vol. 3070, pp. 172–177. Springer (2004)
4. Meeta, K., Anand, K.J., Chandra, S.S.: Socio evolution and learning optimization algorithm: a socio-inspired optimization methodology. Futur. Gener. Comput. Syst. **81**, 252–272 (2018). https://doi.org/10.1016/j.future.2017.10.052
5. Günter, R.: Convergence of evolutionary algorithms in general search spaces. In: Proceedings of 3th IEEE Conference on Evolutionary Computations ICEC, pp. 50–54. IEEE Press (1996)
6. Borovkov, A.A.: Probability Theory. London Ltd., London, GB, Springer (2013)
7. Schaefer, R.: Asymptotic behavior of the artificial genetic systems. In: Foundations of Global Genetic Optimization. Studies in Computational Intelligence, vol. 74. Springer, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73192-4_4
8. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer (1996)
9. Günter, R.: Local convergence rate of similar evolutionary algorithms with cauchy mutations. IEEE Trans. Evol. Comput. **1**(4), 249–258 (1997)
10. Günter, R.: Self-adaptive mutations may lead to premature convergence. IEEE Trans. Evol. Comput. **5**(4), 410–414 (2001)
11. Wróblewski, J.: Adaptive aspects of combining approximation spaces. In: Pal, S.K., Polkowski, L., Skowron, A. (eds.) Rough-Neural Computing. Cognitive Technologies. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-642-18859-6_6
12. Yampolskiy, V.R.: Analysis of types of self-improving software. In: Bieger, J., Goertzel, B., Potapov, A. (eds.) Artificial General Intelligence. AGI 2015. Lecture Notes in Computer Science, vol. 9205. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21365-1_39

13. Nasimul, N., Danushaka, B., Hitoshi, I.: Differential evolution with self adaptive local search. In: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, GECCO'11, pp. 1099–1106 (2011). https://doi.org/10.1145/2001576.2001725
14. Li, W., Sun, Y., Huang, Y., Yi, J.: An adaptive differential evolution algorithm using fitness distance correlation and neighbourhood-based mutation strategy. Connect. Sci. **34**(1), 829–856 (2022). https://doi.org/10.1080/09540091.2021.1997913

Part II
Responsible and Trustworthy Artificial Intelligence

Chapter 12
Explaining and Auditing with "Even-If": Uses for Semi-factual Explanations in AI/ML

Eoin M. Kenny, Weipeng Huang, Saugat Aryal, and Mark T. Keane

Abstract Very recently, semi-factual explanations have emerged in Explainable AI (XAI) as a new and potentially important explanation strategy. Semi-factuals employ "Even if..." reasoning, as opposed to the "If only..." reasoning of counterfactuals. Counterfactuals inform users about what feature-differences lead to changes in an outcome (e.g., "*if only* you asked for a lower loan, you would have been successful."), whereas semi-factuals inform them about what feature-differences lead to the outcome remaining the same (e.g., "*Even if* you asked for a lower loan, you would *still* have been unsuccessful"). Semi-factuals have the potential to be as important as their popular counterfactual siblings. However, the AI/ML and XAI communities have by and large struggled to imagine useful application-scenarios for semi-factuals. In this paper, we summarize recent work on semi-factual explanation and trace a roadmap for application-focused research in the area. We begin by outlining the main constraints identified for semi-factual optimization proposed in the literature, before summarizing the applications of semi-factuals proposed to-date. Then, we sketch several directions for future applications and research using semi-factuals. Finally, though semi-factuals are highly promising (especially with regard to algorithmic recourse), they have a potential for ethical misuse that we discuss in our conclusions.

E. M. Kenny
CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: ekenny@mit.edu

W. Huang
Tencent Security Big Data Lab, Shenzhen, China
e-mail: fuzzyhuang@tencent.com

S. Aryal · M. T. Keane (✉)
School of Computer Science, University College Dublin, Dublin, Ireland
e-mail: mark.keane@ucd.ie

S. Aryal
e-mail: saugat.aryal@ucdconnect.ie

M. T. Keane
Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_12

12.1 Introduction

In Explainable AI (XAI) there are now many post-hoc explanation strategies that have been deployed to explain the decisions of black-box AI systems, from saliency maps [5], to feature-importance methods [39] and instance-based explanations (e.g., factual-case-based explanations [23], counterfactuals [20] and prototypes [26]). In this paper, we focus on a relatively new instance-based strategy that has received much less attention until very recently; namely, the use of semi-factual explanations. Our aim is to highlight the potential of this new example-based XAI strategy and to explore its utility in Artificial Intelligence (AI) and Machine Learning with reference to state-of-the-art research.

Factual, example-based explanations have been used extensively in case-based reasoning (CBR) for decades, where a nearest-neighbour instance is used to explain a prediction made (e.g., "*You were refused this loan, because your profile is the same as a previous customer who was also refused.*"). Counterfactual explanations provide an alternative view of a decision-scenario in which the outcome is shown to change when certain input features change (e.g., "*If you asked for a lower loan-amount, you would have been granted it.*"). Semi-factual explanations, which have been somewhat overlooked, provide a another type of alternative view of the decision-scenario, one in which the outcome is shown *not* to change when certain input-features change (e.g., "*Even if your credit rating was better, you would still have been refused the loan.*"). Semi-factuals have been extensively studied for several decades in psychology [8, 9, 30, 36, 42] and philosophy [6, 7, 14], but have largely passed over in computer science until relatively recently [4, 22, 24, 28, 46, 51, 53].

The fundamental problem that motivates this paper is that, by and large, the AI/ML community has struggled to see how these explanations might be used, in practice [3, 4]. Hence, the aim of the current paper is to surface some of the useful applications of semi-factuals in ML, and to give researchers some clarity on future research directions in this area. Moreover, we also note the potential ethical misuses of this type of explanation. However, we begin by summarizing the methodological requirements for semi-factual explanation (next section) before considering some of the methods that have been proposed to implement it and how they have been applied in different application contexts.

12.2 Constraints on Semi-factuals

Formally, given a test instance x classified as class c, a semi-factual is a modification of x we label x', that remains classified as class c; that is, the changes to x do *not* result in x' crossing a decision boundary into any other class. The more commonly known counterfactual explanation, we label x'', is a modification of x, that changes the classification to another class c'; that is, the changes to x result in x'' crossing a decision boundary into another class. Strictly speaking, any x' that meets the above computational requirement is a semi-factual. However, if we want to use these

semi-factuals as useful explanations, there is general agreement that several other constraints need to be met (see [4] for desiderata that reflect them). Notably, many of these constraints echo those proposed for counterfactual explanations, though they are often realized in different ways:

- *Distance from Test x*. Work on semi-factual explanation has proposed that x' should likely be maximally distant from x [24, 35]. So, if x is "*John was under the legal alcohol limit for driving after he drank 4 units*", the semi-factual x' should say "*Even if John had drunk 7 units, he would still be under the limit*" to make a stronger explanation (rather than e.g. 5 units).
- *Plausibility & Robustness*. However, like counterfactuals, the mutation from x to x' still needs to be plausible [21, 27]. A semi-factual that strays from the known distribution would not make a good explanation. Similarly, as is the case with counterfactuals, the semi-factual needs to be robust [18, 43]. That is, it should not change arbitrarily after small perturbations to the input features (e.g. John drinking 7 units over a duration of 120 mins v. 122 minutes should ideally not change the outcome).
- *Sparsity*. Good semi-factuals explanations should be sparse. That is, they should have few feature-differences to be humanly comprehensible and possibly more convincing [3, 4] (as has been argued for counterfactuals [21, 32, 48]). Indeed, following traditional uses of semi-factuals in the Cognitive Sciences, Aryal and Keane [4] argued they should ideally have just one feature difference. Notably, semi-factuals differ from counterfactuals in that they do not have to cross a decision boundary, so in theory it should only ever be necessary to change one feature; though causal constraints and other practical considerations may make such minimal sparsity uncommon.
- *Distance from the Counterfactual*. Many have proposed that semi-factuals should be positioned between the test x and a good counterfactual for x, x'' [10, 24, 35] (see Fig. 12.1). Kenny and Keane [24] explicitly use the counterfactual, x'', to guide the perturbations made to x when generating the semi-factual x' (specifically, perturbing exceptional features). A weaker version of this claim is that a semi-factual should carefully balance the relative distances to the test and the counterfactual class [4, 10]; specifically, that x' needs to be identifiably in the test's class, c, whilst far enough away from x to be close to the counterfactual class, but not close enough to be perceived as being in the counterfactual class. So, a semi-factual that says "*Even if John had drunk **9** units, he would still be under the limit*" could be a valid semi-factual, but may be less good because it goes too far, it is too close to the decision boundary of known counterfactuals (such as, that "if John drank *10* units, he would definitely be *over* the limit").
- *Diversity*. As has been argued for counterfactuals [32], when several semi-factuals are being generated for a single test instance, they should be diverse (i.e., distinctly different from one another in the feature-differences used) [3].
- *Positive-Outcome Bias*. Finally, very recently, Kenny and Huang [22] have argued that semi-factuals are better at explaining positive outcomes in recourse situations (e.g., a user asking for a higher loan and still being accepted; see Fig. 12.1). They

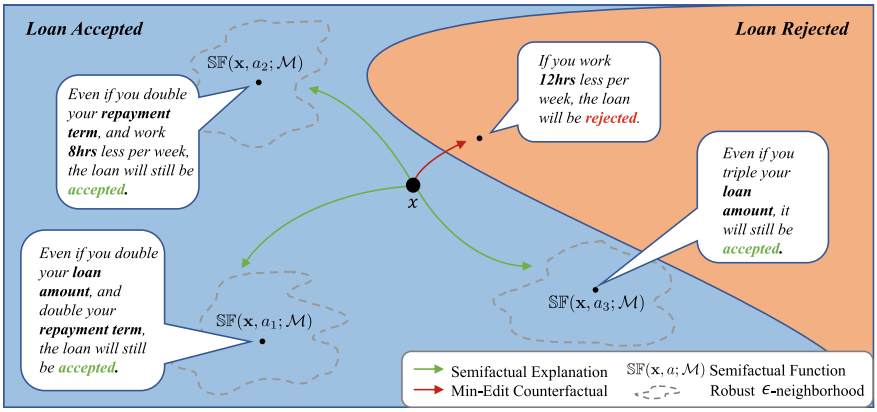


Fig. 12.1 A graphical representation of multiple semi-factuals for a test instance, x, that deliver alternative gains in recourse for a positive outcome (in the loan-accepted class) along with a counterfactual over the decision boundary with a negative outcome (in the load-rejected class). Adapted from Kenny and Huang [22]

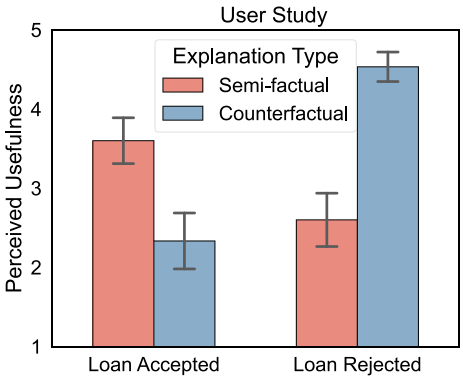


Fig. 12.2 Kenny and Huang's [22] user study results, showing that people rate semi-factual explanations as being significantly more useful than counterfactuals for a positive outcome, whereas the opposite is the case for a negative outcome (bars are standard errors). Adapted from Kenny and Huang [22]

propose a novel "gain" function should be maximized when generating semi-factuals (akin to the idea of *cost* in counterfactual recourse [20]). They also found that people rated semi-factuals to be more useful than counterfactuals in such situations (see Fig. 12.2).

To summarize, many of the constraints used for counterfactuals can be applied to semi-factuals (i.e., sparsity, distance, plausibility, diversity, robustness), albeit sometimes differently. A somewhat naive (but perhaps useful) interpretation of these constraints tells us that counterfactuals (1) *minimize* the featural changes to x, while (2) crossing a decision boundary, while semi-factuals (1) *maximize* the featural changes to x, while (2) *not* crossing a decision boundary. However, semi-factuals also have their own specific constraints (e.g., distance from counterfactuals, positive-outcome bias, gain etc.). Indeed, [22] argued that the application context for semi-factuals

should be quite different to that of counterfactuals, that semi-factuals should be the default explanation-option for positive outcomes, whereas counterfactuals are best used as the default for negative outcomes (a proposal supported by their user study). In the next section, we consider how semi-factual explanations have actually been used in the AI/ML literature, before considering future application contexts.

12.3 Historical Uses of Semi-factuals

Historically, in the 2000s, AI research on semi-factual explanation (not called this at the time) emerged as a type of example-based explanation in CBR, called *a-fortiori* reasoning [10, 12, 34]. These researchers noted that sometimes the most convincing example-based explanation was actually not the nearest neighbour of a test instance, but one that is slightly more distant. Nugent et al. [35] gave the example of trying to convince a child that they shouldn't be allowed go to a scary movie; if the parents used a nearest neighbour to their child (e.g., "*John, who is the same age as you, was not allowed go to the movie*") the argument would be less convincing than using a more distant neighbour (e.g., "*Billy, who is three years older than you, was not allowed go to the movie*"). *A-fortiori reasoning* tries to produce stronger explanations to support some initial proposition.

Doyle et al. [13] deployed a medical application using this type of explanation for a bronchiolitis treatment e-Clinic decision support system with explanations like "*this patient can be released because there is a patient with a much worse condition who was released last week*". The evaluation showed that this type of explanation enhanced the perceived usefulness of the system for doctors. However, this method had the drawback of depending on a hand-coded explanation-utility function to select semi-factuals (e.g., they had to re-calibrate it during their user study as an initial coding diverged from doctor's assessments). Unfortunately, much of this research occurred before extensive computational evaluations were the norm, so the later improved algorithms were often only presented with indicative outputs, rather than rigorous tests; though the datasets explored included ones for blood alcohol, diabetes, and credit card.

12.4 Recent Research on Semi-factuals

In the last 2-3 years, there has been a significant uptick in different applications using semi-factuals following work on a unitary mechanism for generating semi-factuals and counterfactuals [24]. In considering SmartAg applications [25], Kenny and Keane [24] argued that semi-factual explanations could be used to help farmers avoid excess fertiliser use (nitrogen with a high carbon cost that pollutes waterways [44]), using explanations such as "*Even if you double your nitrogen use this month, your crop yield will still be the same.*" (as excess fertiliser just gets washed

out of the soil). This work defines the use-case for semi-factuals in XAI, but it also pointed towards other uses in analysing decision spaces.

This work was quickly followed by papers applying semi-factuals in for auditing use-cases, deploying them to understand class boundaries, feature spaces, spurious features and other uses. Artelt and Hammer [2, 3] proposed that semi-factuals could be applied to elucidate reject decisions, when an ML model refuses to make a prediction due to low predictive certainty. They used semi-factual instances to show why it's reasonable for a prediction *not* to be made; that is, if the semi-factual with higher certainty than query can be rejected, it makes sense for the query to be rejected. In a similar vein, Lu et al. [28] showed how semi-factual augmentations of instances can be used to decouple spurious associations and model bias; they semi-factually generate non-rationale words and re-train the model by replacing original instances to identify spurious features (i.e., feature changes that have absolutely no impact on predictive outcomes). Dandl et al. [11] have extended these ideas to capturing regions in the decision space, what they call "*interpretable region descriptors*", from which semi-factual explanations can be derived. They use these region-descriptors to justify a prediction by providing a set of "even if" arguments (i.e., semi-factual explanations), to indicate which features affect a prediction and whether point-wise biases or implausibilities exist. Finally, Mertes et al. [31] have proposed the novel idea of alterfactuals, a more specialized form of semi-factual that tracks instances that parallel a decision boundary by altering irrelevant features that do *not* contribute to the prediction. Notably, though all of these papers emphasize the auditing potential of semi-factuals they still make use of their explanatory strengths as they try to convince end-users of the decision-space analyses being advanced.

Another set of recent applications has explored the idea of using semi-factuals in different explanatory ways communicating useful information. Zhou [54] has proposed the idea of Iterative Partial Fulfilment which, in essence, uses semi-factuals to explore partial recourses for negative decisions (e.g., when a user may not be able to make all the changes that would prompt the flipping of a decision). Other work [1] has looked at formalizations of semi-factuals as part of a framework to create more personalized explanations to meet user preferences. In representation learning, Vats et al. [46] used generative models for semi-factual explanations in medical domains; specifically, explanations of classifications of images of ulcers. Their novel proposal was that semi-factuals could be used to trace the inflammatory changes in the stomach lining that occur just before ulcer forms, thus providing useful early medical interventions (i.e., tracing the changes from the is-healthy-tissue class to the is-ulcerated-tissue class). Indeed, our own discussions with medical professionals suggest there are a broad set of use-cases for semi-factuals in situations where they want to plot progressive changes in a condition just prior to a significant diagnostic change (e.g., what does a Grade-II or Grade-III muscle tear look like before it becomes a more serious Grade-IV tear). Kenny and Keane's [24] PIECE step-changes exceptional features of the test instance to progressively generate diverse semi-factuals. However, Zhao et al. [53] proposed a class-to-class variational encoder (C2C-VAE) with low computational cost to do the same task applying it to several classic datasets (e.g., MNIST, Yeast, Seeds, Pima, Credit, white wine and so on; see

[52] for Xie et al. [51] use a joint Gaussian mixture model to sample semi-factual images in a deep learner for general model explanation and visualization addressing similar functionalities and Wang et al. [49] have proposed a role for semi-factuals in relation learning for knowledge graphs from text samples. Taken together all of these papers show the generality of applications in which these methods can be useful (albeit mainly for tabular and image datasets). Indeed, practical methodologies involving semi-factuals have also now been proposed for the evaluation ML systems [17] and cognitive tutorials for users [33].

12.5 Future Directions for Semi-factuals

Overall, it is clear that the potential for semi-factual methods is only beginning to be explored. When one looks at counterfactual methods, one sees a bifurcation in application developments between work on XAI for end-users (e.g., the counterfactual XAI literature [15, 47]) and work on ML to analyse decision spaces (e.g., the adversarial learning literature [37, 43]). In reviewing recent semi-factual research, the same bifurcation seems to be emerging between their use in XAI and their use in ML. Both of these directions are clearly areas of future development. In the remainder of this section, we review very recent developments that show fruitful directions for future applications.

Very recent work points to a significant growth-point for semi-factual XAI and algorithmic recourse using the notion of "gain" (analogous to "cost" in counterfactuals [45]). Kenny and Huang [22] have just proposed an innovative semi-factual method, *Semi-Factual Generation* (S-GEN), that develops the notion of "gain" to drive the selection of semi-factual explanations (they also added a new causal analysis [19, 38]). This method hinges on a novel insight about semi-factual use, namely that semi-factuals can give better "recourse" for positive outcomes. For example, if someone has their house-loan application accepted under certain feature conditions (e.g., working 38 hours a week, for a 20-year term), there may be a set of semi-factual changes to features that offer them a better deal (i.e., they "gain" more), such as cases where they could work 8hrs less and/or pay the loan back over a longer 30-year term (see Fig. 12.1).

This work opens up many new avenues for both explainability and the use of semi-factuals for auditing. From the explainability perspective it suggests several new directions. Firstly, in the consideration of positive-outcome advice in other domains, such as warning people off over prescription of medicines [40], where based on medical research, semi-factuals could explain that "*Even if you take half your dose of drug x, you will still be at a low risk of disease y*". Notably, this advice that would benefit patients, rather than corporate entities.

Secondly, it is clear from these new developments, there is much more to be done on ML-auditing uses of semi-factuals. It has been argued psychologically and computationally that semi-factuals weaken the causal connection between target features and outcomes [3, 4, 24, 30]. If you are told "*Jane was under the alcohol*

limit after she drank 4 units" and a valid semi-factual tells you "*Even if Jane had drunk 7 units, she would still be under the limit*", this weakens the causal strength of the units-of-alcohol feature. Indeed, it may suggest that there is another factor at play in the domain (e.g., a latent feature). For example, it may well be that Jane's weight and height combine to process alcohol faster, reducing the impact of the units feature. As such, apart from identifying spurious or redundant features, semi-factuals could guide ML-developers to uncover productive regions of the decision space too.

Thirdly, following Vats et al.'s [46] proposals, semi-factuals could (and perhaps should) be used to generate medical images to aid in timely medical interventions. For example, papillary thyroid carcinoma is now being "over-diagnosed" (and hence treated) globally as the world's fastest growing cancer, particularly in women [16, 41], leading to many unnecessary surgeries reducing quality of life due to the need for life-long thyroid replacement hormones [29]. This cancer is often best treated with active surveillance, but if the tumour grows too big (or too close to the *recurrent laryngeal nerve* needed for speaking) surgical intervention is needed [50]. Semi-factual images could guide doctors in determining the acceptable limit for this cancer's growth to better time surgical intervention. This application of semi-factuals could potentially dramatically improve the quality of life for many people globally.

Finally, the insights on semi-factuals and counterfactuals usage suggest there could be a combined counterfactual and semi-factual logic in a single explanation. For instance, when administering a dosage of a medicine, it may be beneficial to a user to understand the minimum effective dose, but also the maximum effective dose; so, counterfactuals could be used for the former, and semi-factuals for the latter. In this situation, an explanation might say "*If you take x amount of drug y, you will begin to see benefit, but even if you take z amount of drug y, you won't see any additional benefit.*" [1]

12.6 Semi-factuals: Ethical Issues and Dangers

Finally, it is important to be aware of ethical issues and dangers that could arise in the XAI use of semi-factuals. McCloy and Byrne [30] found that counterfactuals lead people to judge the mutated features to be more causally related to the outcome, but semi-factuals had the opposite effect, leading people to judge the mutated features to be less causally related to the outcome. So, semi-factuals weaken causal links between inputs and outcomes, convincing people that the outcome would have occurred anyway. Hence, ethical misuse is possible by leading users to believe that (i) certain features are not important, when in fact the opposite is true [4], and/or (ii) a bad outcome was inevitable, when in fact it may not be [3].

These unethical uses of the method could have many negative impacts on people. Semi-factuals could be used to purposefully teach users incorrect information for

[1] Note, this idea was suggested by Hima Lakkaraju (Harvard University) in discussions with the authors.

malicious purposes, such as trying to get people to take action on certain features in recourse rather than others. They could also be used to reinforce inaction, for example in convincing people to continue doing things that may not be in their best interest (e.g., "*Even if you lose 10lbs and half your smoking, you'll still need to take x amount of our drug each day.*").

Acknowledgements This research was supported by Science Foundation Ireland via the Insight SFI Research Centre for Data Analytics (12/RC/2289). For the purpose of Open Access, the author has applied a CC BY copyright to any Author Accepted Manuscript version arising from this submission.

References

1. Alfano, G., Greco, S., Mandaglio, D., Parisi, F., Shahbazian, R., Trubitsyna, I.: Even-if explanations: formal foundations, priorities and complexity (2024). arXiv:2401.10938
2. Artelt, A., Visser, R., Hammer, B.: "i do not know! but why?âŁž-local model-agnostic example-based explanations of reject. Neurocomputing **558**, 126722 (2023)
3. Artelt, A., Hammer, B.: "even if · · · "-diverse semifactual explanations of reject. In: 2022 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 854–859 (2022)
4. Aryal, S., Keane, M.T.: Even if explanations: prior work, desiderata & benchmarks for semi-factual xai. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, pp. 6526–6535 (Aug 2023). https://doi.org/10.24963/ijcai.2023/732
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7) (2015)
6. Barker, S.: "even, still" and counterfactuals. Linguist. Philos. 1–38 (1991)
7. Bennett, J.: Even if. Linguist. Philos. **5**(3), 403–418 (1982)
8. Boninger, D.S., Gleicher, F., Strathman, A.: Counterfactual thinking: from what might have been to what may be. J. Pers. Soc. Psychol. **67**(2), 297 (1994)
9. Byrne, R.M.: Precis of the rational imagination: how people create alternatives to reality. Behav. Brain Sci. **30**(5–6), 439–453 (2007)
10. Cummins, L., Bridge, D.: Kleor: a knowledge lite approach to explanation oriented retrieval. Comput. Inf. **25**(2–3), 173–193 (2006)
11. Dandl, S., Casalicchio, G., Bischl, B., Bothmann, L.: Interpretable regional descriptors: hyperbox-based local explanations (2023)
12. Doyle, D., Cunningham, P., Bridge, D., Rahman, Y.: Explanation oriented retrieval. In: European Conference on Case-Based Reasoning, pp. 157–168. Springer (2004)
13. Doyle, D., Cunningham, P., Walsh, P.: An evaluation of the usefulness of explanation in a case-based reasoning system for decision support in bronchiolitis treatment. Comput. Intell. **22**(3–4), 269–281 (2006)
14. Goodman, N.: Fact, fiction, and forecast. Harvard University Press (1983)
15. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Min. Knowl. Discov. 1–55 (2022)
16. Hoang, J.K., Nguyen, X.V., Davies, L.: Overdiagnosis of thyroid cancer: answers to five key questions. Acad. Radiol. **22**(8), 1024–1029 (2015)
17. Hoffman, R.R., Jalaeian, M., Tate, C., Klein, G., Mueller, S.T.: Evaluating machine-generated explanations: a "scorecard" method for xai measurement science. Front. Comput. Sci. **5**, 1114806 (2023)

18. Jiang, J., Leofante, F., Rago, A., Toni, F.: Formalising the robustness of counterfactual explanations for neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14901–14909 (2023)
19. Karimi, A.H., Barthe, G., Belle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions (2019). arXiv:1905.11190
20. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: definitions, formulations, solutions, and prospects (2021)
21. Keane, M.T., Smyth, B.: Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable ai (xai). In: International Conference on Case-Based Reasoning. Springer (2020)
22. Kenny, E.M., Huang, W.: The utility of "even if" semi-factual explanation to optimize positive outcomes. In: NeurIPS-23 (2023)
23. Kenny, E.M., Keane, M.T.: Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai. In: Twenty-Eighth International Joint Conferences on Artificial Intelligence (IJCAI), Macao, 10–16 Aug 2019, pp. 2708–2715 (2019)
24. Kenny, E.M., Keane, M.T.: On generating plausible counterfactual and semi-factual explanations for deep learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11575–11585 (2021)
25. Kenny, E.M., Ruelle, E., Geoghegan, A., Shalloo, L., O'Leary, M., O'Donovan, M., Temraz, M., Keane, M.T.: Bayesian case-exclusion and personalized explanations for sustainable dairy farming. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (2020)
26. Kenny, E.M., Tucker, M., Shah, J.: Towards interpretable deep reinforcement learning with human-friendly prototypes. In: The Eleventh International Conference on Learning Representations (2023)
27. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations (2019). arXiv:1907.09294
28. Lu, J., Yang, L., Namee, B., Zhang, Y.: A rationale-centric framework for human-in-the-loop machine learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1. Long Papers, pp. 6986–6996 (2022)
29. Massolt, E.T., van der Windt, M., Korevaar, T.I., Kam, B.L., Burger, J., Franssen, G.J., Lehmphul, I., Köhrle, J., Visser, W.E., Peeters, R.P.: Thyroid hormone and its metabolites in relation to quality of life in patients treated for differentiated thyroid cancer. Clin. Endocr. **85**(5), 781–788 (2016)
30. McCloy, R., Byrne, R.M.: Semifactual "even if" thinking. Think. Reason. **8**(1), 41–67 (2002)
31. Mertes, S., Karle, C., Huber, T., Weitz, K., Schlagowski, R., André, E.: Alterfactual explanations–the relevance of irrelevance for explaining ai systems. IJCAI-22 Workshop on XAI (2022)
32. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 607–617 (2020)
33. Mueller, S., Tan, Y.Y., Linja, A., Klein, G., Hoffman, R.: Authoring guide for cognitive tutorials for artificial intelligence: purposes and methods (2021)
34. Nugent, C., Cunningham, P., Doyle, D.: The best way to instil confidence is by being right. In: International Conference on Case-Based Reasoning, pp. 368–381. Springer (2005)
35. Nugent, C., Doyle, D., Cunningham, P.: Gaining insight through case-based explanation. J. Intell. Inf. Syst. **32**(3), 267–295 (2009)
36. Parkinson, M., Byrne, R.: Counterfactual & semi-factual thoughts in moral judgements about failed attempts to harm. Think. Reason. **23**, 409–448 (2017)
37. Pawelczyk, M., Agarwal, C., Joshi, S., Upadhyay, S., Lakkaraju, H.: Exploring counterfactual explanations through the lens of adversarial examples: a theoretical and empirical analysis. In: International Conference on Artificial Intelligence and Statistics, pp. 4574–4594. PMLR (2022)

38. Pearl, J.: Causality: models, reasoning and inference, vol. 9, pp. 10–11. Cambridge University Press, Cambridge, MA, USA (2000)
39. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
40. Safer, D.J.: Overprescribed medications for us adults: four major examples. J. Clin. Med. Res. **11**(9), 617 (2019)
41. Sanabria, A., Kowalski, L.P., Shah, J.P., Nixon, I.J., Angelos, P., Williams, M.D., Rinaldo, A., Ferlito, A.: Growing incidence of thyroid carcinoma in recent years: factors underlying overdiagnosis. Head & Neck **40**(4), 855–866 (2018)
42. Sarasvathy, S.D.: Even-if: sufficient, yet unnecessary conditions for worldmaking. Organization Theory **2**(2), 26317877211005784 (2021)
43. Slack, D., Hilgard, A., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. Adv. Neural Inf. Process. Syst. **34**, 62–75 (2021)
44. Suddick, E.C., Whitney, P., Townsend, A.R., Davidson, E.A.: The role of nitrogen in climate change and the impacts of nitrogen-climate interactions in the united states: foreword to thematic issue. Biogeochemistry **114**, 1–10 (2013)
45. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 10–19 (2019)
46. Vats, A., Mohammed, A., Pedersen, M., Wiratunga, N.: This changes to that: combining causal and non-causal explanations to generate disease progression in capsule endoscopy (2022). arXiv:2212.02506
47. Verma, S., Dickerson, J.P., Hines, K.: Counterfactual explanations for machine learning: a review (2022). CoRR arxiv:abs/2010.10596
48. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
49. Wang, J., Zhang, L., Liu, J., Guo, T., Wu, W.: Learning from semi-factuals: a debiased and semantic-aware framework for generalized relation discovery (2024). arXiv:2401.06327
50. Wu, C.W., Dionigi, G., Barczynski, M., Chiang, F.Y., Dralle, H., Schneider, R., Al-Quaryshi, Z., Angelos, P., et al.: International neuromonitoring study group guidelines 2018. Laryngoscope **128**, S18–S27 (2018)
51. Xie, Z., He, T., Tian, S., Fu, Y., Zhou, J., Chen, D.: Joint gaussian mixture model for versatile deep visual model explanation. Knowl. Based Syst. (2023)
52. Ye, X., Leake, D., Wang, Y., Zhao, Z., Crandall, D.J.: Selecting feature changes for counterfactual explanation. In: IJCAI-22 Workshop on XAI (2022)
53. Zhao, Z., Leake, D., Ye, X., Crandall, D.: Generating counterfactual images: towards a c2c-vae approach. In: 4th Workshop on XCBR: Case-Based Reasoning for the Explanation of Intelligent Systems (2022)
54. Zhou, Y.: Iterative partial fulfillment of counterfactual explanations: benefits and risks. In: AIES'23, pp. 248–258 (2023)

Chapter 13
Automatic Classification and Localization of Ancient Amphorae Through Object Detection in Underwater Archeology

Lucia Lombardi, Francesco Mercaldo, and Antonella Santone

Abstract Underwater archaeology involves the systematic documentation and recovery of information from submerged artifacts, such as pottery amphorae, at underwater sites. Computer vision can significantly aid underwater archaeologists by enabling the automatic detection of these artifacts from images taken underwater by both archaeologists and robots. In this paper, we propose a method for the automatic detection and localization of ancient amphorae in underwater images. Our experimental analysis, conducted on an annotated dataset containing over 1,600 images, demonstrated the effectiveness of the proposed method.

13.1 Introduction and Related Work

Sea, nature, and culture: this is an indissoluble trinomial when it comes to underwater archeology. Underwater archeology is a specialized branch of archeology that focuses on the study and documentation of submerged archeological sites and artifacts. Underwater archeology is about the recovery of information from submerged artifacts, for instance, pottery amphorae, from underwater sites.

Bringing to light fragments of ancient artifacts, for instance, amphorae is a very important way to learn about our past. The old ships that are often found off the coasts preserve invaluable information on the civilizations that preceded us, helping us to understand numerous phenomena relating to trade, wars, and migrations of our ancestors. Yet, probing the surface of the seabed is often complicated and expensive, as well as dangerous.

L. Lombardi · F. Mercaldo (✉) · A. Santone
Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, Campobasso, Italy
e-mail: francesco.mercaldo@unimol.it

L. Lombardi
e-mail: l.lombardi12@studenti.unimol.it

A. Santone
e-mail: antonella.santone@unimol.it
© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_13

Wrecks constitute the classic case to be investigated in underwater archeology, but prospecting and excavation interventions are carried out not only on naval wrecks but also on submerged structures, such as the amphora. As a matter of fact, the amphora is perhaps the most emblematic of the finds found at the bottom of the sea.

Evidence of this is both a series of wrecks discovered and investigated throughout the Mediterranean with loads of amphorae and the excavations conducted in all the cities of the ancient world which have returned large quantities of ceramic fragments, attributable to transport amphorae, coming from all over the Mediterranean basin.

Amphorae plays a fundamental role in comprehending archeological contexts, offering insights into production, trade networks, and social interactions. However, amphorae discovering remains primarily a manual process, reliant on analog catalogs curated by archeologists and stored in archives and libraries.

Computer vision can provide great support for underwater archeologists by providing support for the automatic detection of submerged artifacts from images acquired underwater by archeologists and robots. Unfortunately, underwater archeology often turns out to be expensive and dangerous. However, the widely varying underwater conditions pose challenges in achieving high-performance results.

Recent studies demonstrated the possibility of detecting objects in underwater environments by exploiting computer vision [14], for instance, authors in [13] proposed an automated system for wreck detection, utilizing the YOLO ("You Only Look Once") 3 deep learning architecture, while researchers in [1] exploited the YOLO 5 model for object detection in underwater environments; specifically, to detect defects in fish farming nets.

Despite numerous research studies, concerning the classification of archeological findings and numerous examples of object detection in underwater conditions, there is a limited amount of research that combines the detection of archeological findings in underwater environments [2, 6, 7]. For these reasons, in this paper, we investigate the possibility of detecting and localizing ancient amphorae in images acquired underwater by exploiting deep learning, in particular object detection. We resort to the YOLO 8 model for object detection, considering its better detection capabilities in comparison to the previous versions, as shown in [5].

The paper proceeds as follows: in the next section we present and describe the proposed method, the results of the experimental evaluation are shown in Sect. 13.3, and, finally, in the last section conclusions and future research lines are drawn.

13.2 The Method

In this section, we present our proposed method for detecting and localizing ancient amphorae in underwater images.

Specifically, we introduce a technique designed to automatically identify ancient amphorae directly from underwater images, which may be captured by robots. Additionally, our method can pinpoint the location of amphora fragments within the image and provide the prediction percentage for each detected fragment.

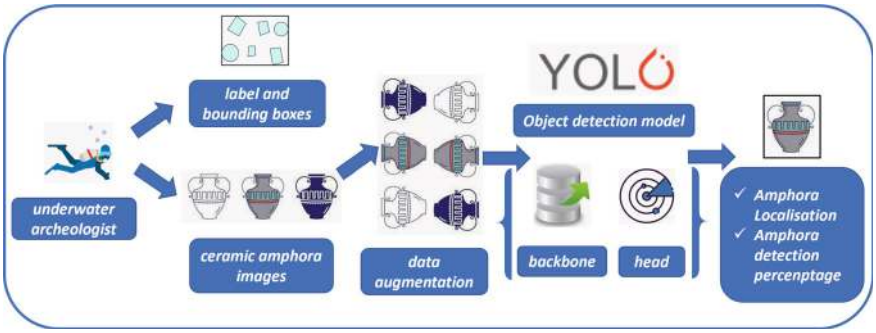


Fig. 13.1 The proposed method

Figure 13.1 depicts the proposed method.

To develop an effective deep learning model for detecting ancient amphorae, it is crucial to have a dataset consisting of images captured by underwater archaeologists or robots (*underwater archeologist* in Fig. 13.1): as a matter of fact, a robot can be particularly useful for acquiring images, as it can access and capture visuals from highly challenging and inaccessible underwater locations. To develop a model capable of not only detecting the presence of ancient pottery amphorae in an image but also pinpointing the location of amphora fragments within the image, a collection of images with detailed information on the amphora's position is necessary. Therefore, these images must be labeled and annotated by domain experts (i.e., *underwater archeologist* in Fig. 13.1) with the aim to mark the area(s) of the images under analysis where the amphora fragment are present (i.e., *label and bounding boxes* in Fig. 13.1). The class for the detection of the bounding box is only one i.e., amphora (i.e., the proposed model is devoted to detect the presence of fragments of amphora).

Additionally, to develop a model that is both effective and capable of accurately predicting unseen images, it is crucial to capture a diverse set of images from various angles and under different conditions. Although the images may initially vary in size, a preprocessing step is necessary to resize them to a uniform dimension.

The subsequent phase focuses on augmenting the number of images, as illustrated by the *image augmentation* in Fig. 13.1. To achieve this, we employ various techniques to expand the existing dataset without additional data collection. Data augmentation involves making controlled random modifications to the existing images, creating altered duplicates. This technique is advantageous for the automated learning process of artificial neural networks, improving their accuracy as the training dataset grows.

Specifically, we utilize data augmentation techniques to create images with controlled random modifications, such as flips [11]. The rationale for applying data augmentation in this context is to ensure the model's effectiveness in accurately recognizing amphora fragments, regardless of their location within the image. Additionally, augmented data helps address the issue of overfitting, which occurs when the model becomes overly tailored to the training data sample.

After acquiring the (augmented) images, along with relevant details about the amphora class and bounding boxes, the next requirement is a deep learning model (i.e., *Object Detection model* in Fig. 13.1).

In this paper, we resort to the YOLO 8 model [8].

Unlike existing object detection models, YOLO demonstrates significantly faster performance, as evidenced in [9, 10].

We choose this model because, compared to alternative deep learning models for object detection, YOLO, despite being recognized for potentially more localization errors [4], shows a reduced tendency to identify false positives in the image background. Additionally, it is notably faster [3, 5]. These characteristics together establish YOLO as widely acknowledged as one of the most effective convolutional neural network models for object detection.

The YOLO network consists of a backbone, which is a convolutional neural network responsible for gathering and organizing image features across different scales, and a Head that utilizes these features from the backbone to perform box and class prediction tasks. Positioned between the backbone and the head is the neck, a series of layers designed to blend and integrate image features before forwarding them to the prediction phase. Specifically, in this study, we investigate the small version of the YOLO 8 model, referred to as YOLOv8s [12].

13.3 Experimental Analysis

In this section, we present the results of our experimental analysis to demonstrate the effectiveness of the YOLO 8 model in detecting amphora fragments from images captured underwater.

We gathered images from the *archeologia Computer Vision Project*, a dataset aimed to build models to detect pottery amphoras in images acquired underwater. The dataset is freely available for research purposes.[1]

The utilized dataset consists of 1699 distinct real-world images featuring ancient amphora fragments. Each image is labeled with a single category, "amphora fragment" (abbreviated as "amphora-frag"), accompanied by its corresponding bounding box indicating the fragment's position within the image. We have chosen to use a publicly available dataset to ensure the reproducibility of our results.

The images of amphora fragments are saved in JPEG format with a resolution of 640 $\times$ 640 pixels. We divided the images as follows: 1431 images for training, 179 for validation, and the remaining 89 for the test set, resulting in a split ratio of 70:20:10.

The dataset we acquired is annotated, meaning each image includes detailed information about the bounding box around each amphora fragment.

We conducted image augmentation by generating additional images for each original image using horizontal and vertical flips.

[1] https://universe.roboflow.com/prova-bjifx/archeologia.

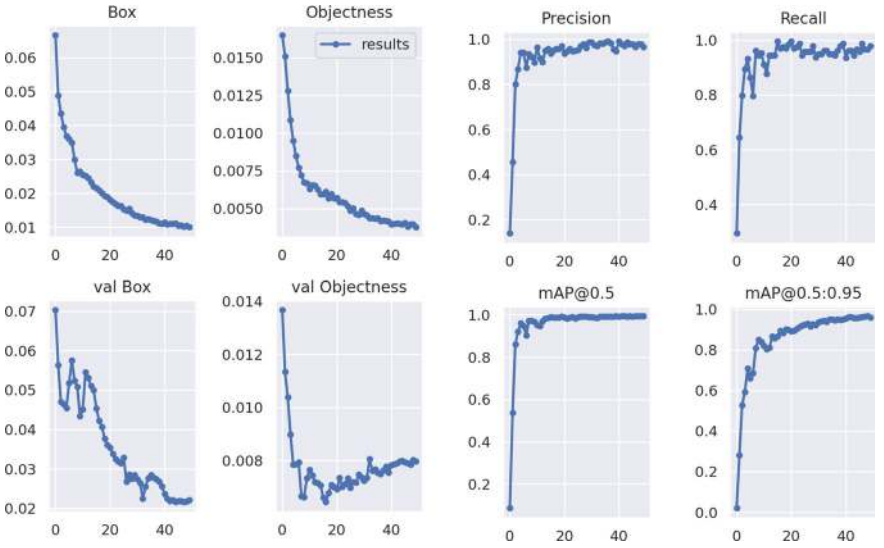


Fig. 13.2 Experimental analysis results

After completing the data augmentation process, we finalized the dataset. All images were resized to dimensions of 640 × 640. For model configuration, a batch size of 16 was selected, and training was conducted over 50 epochs, starting with a learning rate of 0.01.

For model training, we utilized a machine equipped with an NVIDIA Tesla T4 GPU card featuring 16 GB of memory. The model training took 2.255 h (this value refers to the wall clock time). The experimental results achieved by the proposed method are depicted in Fig. 13.2 through various plots. In the first row of plots in Fig. 13.2, the depicted metrics include: box (showing the trend of box_loss during training, which measures how accurately the predicted bounding boxes match the ground truth objects), objectness (showing the trend of object_loss during training), precision trend, and recall trend. In the second row of plots in Fig. 13.2, the displayed metrics include: val Box (showing the trend of box_loss during validation), val Objectness (indicating the trend of object_loss during validation), mean Average Precision with Intersection over Union equal to 0.5 (mAP_0.5), and mean Average Precision with Intersection over Union between 0.5 and 0.95 (mAP_0.5:0.95).

All metrics indicate that the trained model successfully classifies and localizes ancient amphorae. Precision, recall, mAP_0.5, and mAP_0.5:0.95 are expected to show an upward trend as the number of epochs increases, reflecting the model's improved ability to detect objects in underwater images. Conversely, other metrics exhibit a downward trend with increasing epochs, providing additional evidence of the model's effective learning from underwater images. Loss metrics, in particular, highlight instances where the model incorrectly identifies specific objects. Therefore, it is common for loss values to start high in early epochs but decrease over time as the model improves its detection accuracy.

Below, we provide additional information on precision, recall, mAP_0.5, and mAP_0.5:0.95 metrics.

Precision represents the ratio of correctly predicted positive instances to the total predicted positive instances. It accounts for false positives, which are instances incorrectly identified as positive.

Examining the precision trend depicted in Fig. 13.2, the trend clearly shows a rise across multiple epochs until it stabilizes, particularly around the 10th epoch. This pattern indicates that the model gradually improves its ability to distinguish between different ancient amphora from underwater images as training progresses.

The second metric used to evaluate the effectiveness of the proposed method is recall. This metric measures the proportion of actual positives that were correctly predicted by the model. It considers false negatives, which are instances that should have been detected but were missed, offering a thorough assessment of the model's capability to capture all pertinent instances.

Observing the recall trend presented in Fig. 13.2, a behavior akin to the one highlighted for precision becomes apparent. As anticipated, both precision and recall should exhibit an increasing trend with the growing number of epochs. The plots in Fig. 13.2 confirm this upward trend for both metrics, and given that precision and recall values range from 0 to 1, the achieved performances are promising. Much like precision, recall demonstrates an increasing trend as the number of epochs increases.

Precision and recall are widely used metrics for evaluating a model's performance in classification tasks. However, to assess whether the model accurately identifies the object of interest within the correct part of the underwater image being analyzed, additional metrics are necessary. One such metric is Average Precision (AP), a standard measure used to evaluate the accuracy of object detectors like the YOLO 8 model we utilized. Average precision computes the average precision score across recall values ranging from 0 to 1.

We focus on computing the mean Average Precision (mAP), a metric that incorporates Intersection over Union (IOU), Precision, Recall, Precision-Recall Curve, and AP. Object detection models make predictions that include bounding boxes and corresponding object categories within an image. IOU is utilized to evaluate the accuracy of these bounding box predictions.

IOU measures the degree of overlap between bounding boxes, ranging from 1.0 for a perfect match to 0.0 for no overlap.

When assessing object detection models, it's crucial to determine the threshold for bounding box overlap that constitutes successful recognition relative to ground truth data. This is achieved using Intersection over Union (IOU), where mAP_0.5 represents accuracy when IOU is 0.5, indicating successful detection if overlap exceeds 50%. A higher IOU indicates a stricter requirement for accurate bounding box detection, presenting greater difficulty. For instance, mAP_0.75 typically yields a lower value compared to mAP_0.5.

The mAP is an average of the AP values, which is a further average of the APs for all classes.

13.3.1 Results

In Fig. 13.2 are shown, respectively in the metrics/mAP_0.5 and the metrics/mAP_0.5:0.95 plots the mAP value for IOU = 50 and IOU ranging from 50 and 95 (i.e., this value represents different IoU thresholds from 0.5 to 0.95, with a step size equal to 0.05) on average mAP.).

We note that the trends depicted in the plots for the metrics mAP_0.5 and mAP_0.5:0.95 in Fig. 13.2 both exhibit an upward trajectory. The model demonstrates an ability to learn the specific region of the image where attention should be directed to accurately identify the objects to be detected.

Furthermore, to better evaluate the proposed method, in Fig. 13.3 we report the precision and recall values on the Precision-Recall graph.

The expected behavior for this plot is a monotonically decreasing trend. This is due to the inherent trade-off between precision and recall, where increasing one will inevitably decrease the other. While exceptions or data limitations may sometimes disrupt the typical monotonically decreasing pattern in precision-recall graphs, Fig. 13.3 illustrates a decreasing trend for the relevant labels. The precision-recall plot also displays the Area Under the Curve (AUC) values for the `amphora-frag` class and the identification of all classes with mAP@0.5.

Given that these metrics range from 0 to 1, these values can be deemed satisfactory.

Figure 13.1 shows the values obtained for Precision, Recall, mAP_0.5, and mAP_0.5:0.95 metrics (i.e., the average value of the metrics for all the classes involved.).

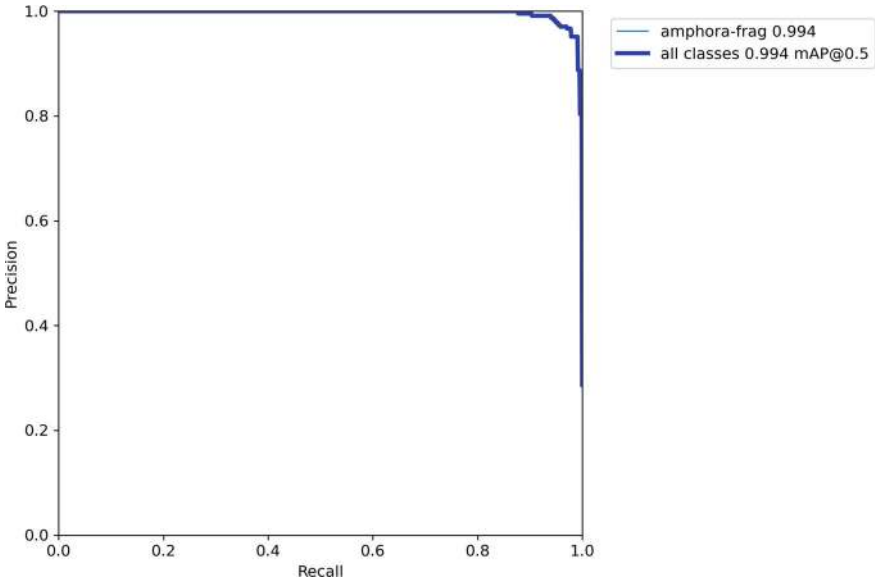


Fig. 13.3 The Precision-Recall graph

From Table 13.1 we can note that the Precision and the Recall are respectively equal to 0.967 and 0.979. Figure 13.4 shows several examples of detection and localization performed by the proposed method.

Table 13.1 Classification results

Precision	Recall	mAP_0.5	mAP_0.5:0.95
0.967	0.979	0.994	0.959

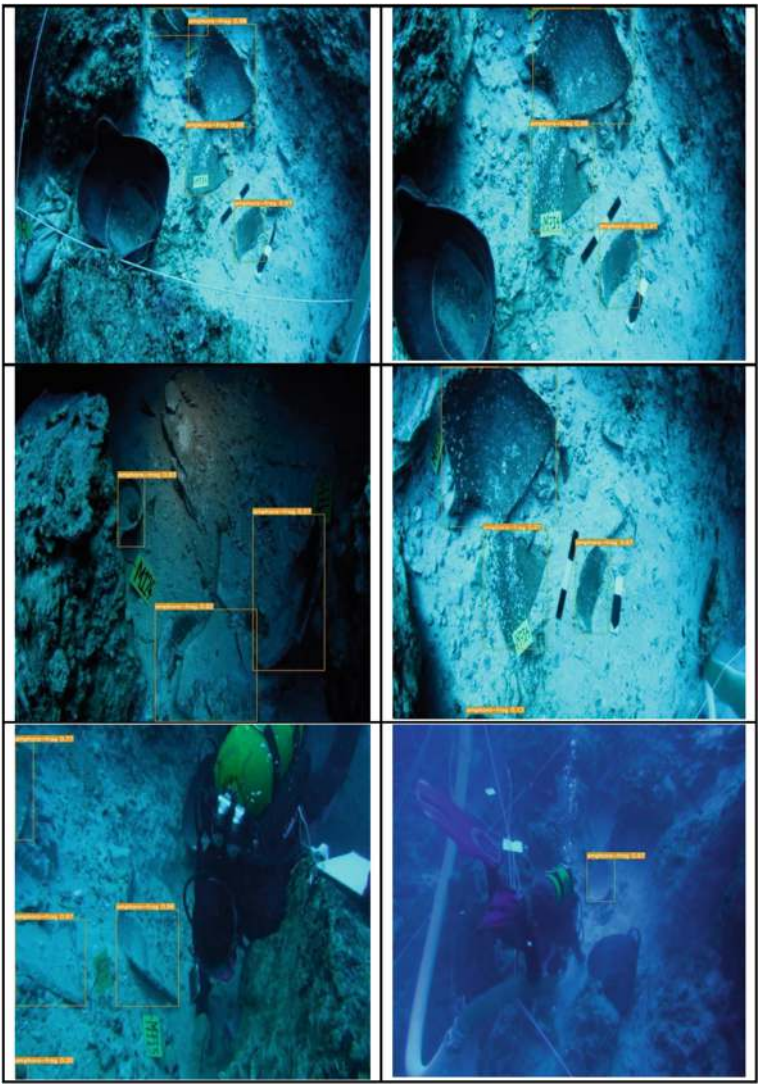


Fig. 13.4 Examples of pottery amphora fragment detection and localization

As shown in Fig. 13.4 the proposed method is able to rightly localize the pottery amphora fragments even if they are confused with the seabed and also in frames where there is the presence of the underwater archeologist, thus confirming the effectiveness of the proposed model.

13.4 Conclusion and Future Work

Given the significant importance of recovering submerged artifacts in underwater archaeology, this paper introduces a method for automatically detecting and localizing ancient amphora fragments from images captured underwater. We employed the YOLO 8 model and a dataset comprising 1699 underwater images. Achieving a precision of 0.967 and a recall of 0.979 demonstrates the effectiveness of our approach. In future work, we intend to explore different versions of the YOLO 8 model, such as YOLOv8 Large (YOLOv8l) and YOLOv8 Extra Large (YOLOv8x), aiming to enhance performance. Additionally, our focus will be on fragment segmentation to obtain more detailed annotations.

Acknowledgements This work has been partially supported by EU DUCA, EU CyberSecPro, SYNAPSE, PTR 22-24 P2.01 (Cybersecurity) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the EU–NextGenerationEU projects, by MUR–REASONING: foRmal mEthods for computAtional analySis for diagnOsis and progNosis in imagING–PRIN, e-DAI (Digital ecosystem for integrated analysis of heterogeneous health data related to high-impact diseases: innovative model of care and research), Health Operational Plan, FSC 2014–2020, PRIN-MUR-Ministry of Health, the National Plan for NRRP Complementary Investments D^3 4 Health: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care, Progetto MolisCTe, Ministero delle Imprese e del Made in Italy, Italy, CUP: D33B22000060001 and FORESEEN: FORmal mEthodS for attack dEtEction in autonomous drIviNg systems CUP N.P2022WYAEW and by Fondazione Intesa SanPaolo Onlus in the "Doctorates in Humanities Disciplines" for the "Artificial Intelligence for the Analysis of Archaeological Finds" topic.

References

1. Al Muksit, A., Hasan, F., Emon, M.F.H.B., Haque, M.R., Anwary, A.R., Shatabda, S.: Yolo-fish: a robust fish detection model to detect fish in realistic underwater environment. Ecolog. Inf. **72**, 101, 847 (2022)
2. Cerrillo-Cuenca, E., de Sanjosé Blasco, J.J., Belinchón, R.C., Bueno-Ramírez, P., Cordero, A.G., Pérez-Álvarez, J.A.: Surveying and monitoring submerged archaeological sites in inland waters through a multiproxy strategy: the case of dolmen de guadalperal and other sites from valdecañas reservoir (Spain). Archaeological Prospection (2024)
3. Horak, K., Sablatnig, R.: Deep learning concepts and datasets for image recognition: overview 2019. In: Eleventh International Conference on Digital Image Processing (ICDIP 2019), vol. 11179, pp. 484–491. SPIE (2019)
4. Hurtik, P., Molek, V., Hula, J., Vajgl, M., Vlasanek, P., Nejezchleba, T.: Poly-yolo: higher speed, more precise detection and instance segmentation for yolov3. Neural Comput. Appl. **34**(10), 8275–8290 (2022)

5. Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B.: A review of yolo algorithm developments. Proc. Comput. Sci. **199**, 1066–1073 (2022)
6. Nigam, R.: Geological/paleontological applications in marine archeology: few examples from Indian waters. In: The Role of Tropics in Climate Change, pp. 419–437. Elsevier (2024)
7. Pydyn, A., Popek, M., Janowski, Ł., Kowalczyk, A., Żuk, L.: Between water and land: connecting and comparing underwater, terrestrial and airborne remote-sensing techniques. J. Archaeol. Sci. Rep. **53**, 104, 386 (2024)
8. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
9. Sah, S., Shringi, A., Ptucha, R., Burry, A.M., Loce, R.P.: Video redaction: a survey and comparison of enabling technologies. J. Electron. Imaging **26**(5), 051, 406 (2017)
10. Sanchez, S., Romero, H., Morales, A.: A review: comparison of performance metrics of pretrained models for object detection using the tensorflow framework. In: IOP Conference Series: Materials Science and Engineering, vol. 844, p. 012024. IOP Publishing (2020)
11. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
12. Yan, T., Sun, W., Cui, K.: Real-time ship object detection with yolor. In: Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning, pp. 203–210 (2022)
13. Yulin, T., Jin, S., Bian, G., Zhang, Y.: Shipwreck target recognition in side-scan sonar images by improved yolov3 model based on transfer learning. IEEE Access **8**, 173,450–173,460 (2020)
14. Zhou, X., Tang, C., Huang, P., Mercaldo, F., Santone, A., Shao, Y.: Lpcanet: classification of laryngeal cancer histopathological images using a cnn with position attention and channel attention mechanisms. Interdiscip. Sci. Comput. Life Sci. **13**(4), 666–682 (2021)

Chapter 14
Exploring Stroke Factors Using Approximate Inverse Model Explanations (AIME): A Method for Extracting Relevant Factors from a Stroke Dataset

Takafumi Nakanishi

Abstract This study proposes a novel approach that applies approximate inverse model explanations (AIME) on a stroke dataset to evaluate the factors that precipitate or prevent stroke occurrence. AIME helps explain the behavior of complex or less transparent AI and machine learning models (black-box models) and the basis of data instance estimates by constructing approximate inverse operators. Unlike previous methods, AIME yields highly interpretable explanations, thereby enhancing the transparency of complex AI and machine learning models for medical diagnosis. By employing AIME to construct an approximate inverse operator from a machine learning black-box model trained on a stroke dataset, this study aims to elucidate factors that may contribute to or mitigate the risk of stroke. Addressing this critical research gap, this method helps elucidate opaque black-box model decisions, potentially revolutionizing stroke risk assessment and prevention strategies. Future studies will aim to extend these interpretations into broader medical applications, foster discoveries, and improve patient outcomes through explainable AI. The potential scalability and adaptability of the proposed method suggest a promising future for medical AI, heralding a new era of explainable and dependable machine learning in healthcare.

14.1 Introduction

Stroke is a serious health problem characterized by sudden onset and enduring consequences that profoundly affect individuals, families, and society at large. Effective prevention, early diagnosis, and appropriate treatment strategies are key in mitigating the impact of stroke. Furthermore, an accurate understanding and prediction of

T. Nakanishi (✉)
Musashino University, Koto-ku, Tokyo 135-8181, Japan
e-mail: takafumi.nakanishi@ds.musashino-u.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_14

stroke-inducing factors are essential. In this context, AI and machine learning models hold significant promise in discerning meaningful patterns from extensive medical datasets to evaluate the likelihood of stroke. However, for these models to gain widespread acceptance and trust within medical practice, the rationale behind their decisions must be transparent and readily comprehensible. Explainable AI (XAI) [1] emerges as a vital component in this endeavor because it facilitates the interpretability of decision-making processes within AI and machine learning models, empowering physicians and patients to comprehend and place confidence in the recommendations provided by AI systems. Such transparency can facilitate the adoption of AI into medical practice, thereby contributing to the effective prevention and treatment of strokes.

To address this challenge, we propose an XAI method, called approximate inverse model explanations (AIME) [2], aimed at deriving approximate inverse operators from complex or less transparent AI and machine learning models (black-box models). Through this derivation, AIME can explain the behavior of the black-box model and the rationale behind its estimations when presented with specific data instances. By building machine learning models for stroke data and using AIME, insights into the primary factors contributing to stroke occurrence or the criteria for determining whether an individual patient has experienced a stroke can be derived. This approach involves pretraining a machine learning model using stroke data and subsequently deriving an approximate inverse operator using AIME. We further hypothesize that these findings will enhance the accuracy of medical diagnoses by ensuring transparency in the integration of AI and machine learning models.

The remainder of this paper is structured as follows: Sect. 14.2 discusses related studies, Sect. 14.3 outlines the proposed method in detail, Sect. 14.4 describes the conducted experiments, and Sect. 14.5 summarizes the results obtained.

14.2 Related Works

In this section, we examine several relevant studies that incorporate XAI techniques to elucidate stroke diagnoses, highlighting the superior efficacy of our model compared with these approaches. Islam et al. [3] developed a machine learning model for stroke estimation utilizing electroencephalography (EEG) data. They employed Local Interpretable Model-agnostic Explanations (LIME) [4] and Explain Like I'm Five (Eli5), a library that helps visualize and understand data and predictions [5], to extract underlying explanations. Gandolfi et al. [6] investigated the feasibility of using machine learning (ML) to accurately predict upper limb (UL) recovery in subacute patients. They employed random forests to isolate the contributions of variables shaping the results, interpreted the outcomes, and evaluated the relevance of features. Although not directly related to XAI, Bhattacharya et al. [7] applied the label encoder technique to stroke data, mitigating data imbalances through oversampling and employing the antlion optimization algorithm for efficient hyperparameter selection in a deep neural network model. This resulted in notable reductions in the model training time. The

novelty of our study lies in its focus on stroke data [8] and the application of the previously studied AIME method [2] to this domain.

14.3 Method

In this section, we outline a methodology for deriving approximate inverse operators from a black-box model to predict stroke occurrence, leveraging AIME to elucidate the model's behavior and the rationale behind its estimations. Section 14.3.1 provides an overview of AIME. Section 14.3.2 describes the global feature importance, which elucidates the model's behavior by utilizing AIME. Section 14.3.3 discusses the local feature importance, which provides the derivation basis for each dataset using AIME for explanation purposes.

14.3.1 Overview of AIME

Figure 14.1 provides an overview of AIME. A detailed explanation of AIME is given in a previous study [2]. In Fig. 14.1, X represents the explanatory variable, which is a matrix of the number of data instances $\times$ number of features, and Y represents the objective variable, which is a matrix of the number of data entries $\times$ number of classes. By training a black-box model on these data and subsequently

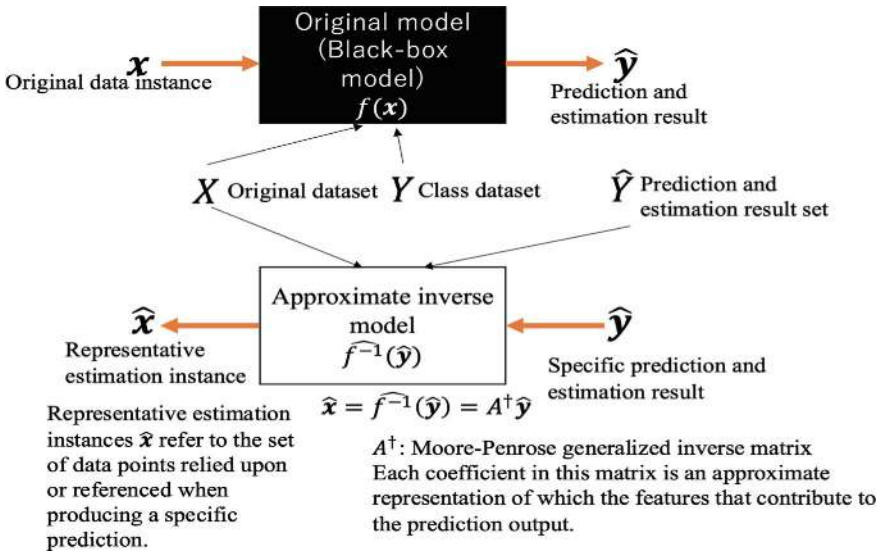


Fig. 14.1 Overview of AIME

inputting a new data instance $\boldsymbol{x}$, an estimate $\hat{\boldsymbol{y}}$ may be derived. Here, the matrix X is decomposed into individual data instances and fed into the black-box model. The outputs are collected in a matrix $\widehat{Y}$, which is a matrix of the number of data instances x number of classes. This $X \rightarrow \widehat{Y}$ function represents the input–output relationship of the black-box model. Incorporating these data, an approximate inverse operator capable of deriving explanations $A^{\dagger}$ can be constructed from the following equation [2]:

$$\begin{aligned} X &= A^{\dagger}\widehat{Y} \\ X\widehat{Y}^{T} &= A^{\dagger}\widehat{Y}\widehat{Y}^{T} \\ X\widehat{Y}^{T}\left(\widehat{Y}\widehat{Y}^{T}\right)^{-1} &= A^{\dagger}\left(\widehat{Y}\widehat{Y}^{T}\right)\left(\widehat{Y}\widehat{Y}^{T}\right)^{-1} \\ A^{\dagger} &= X\widehat{Y}^{T}\left(\widehat{Y}\widehat{Y}^{T}\right)^{-1} = X\widehat{Y}^{\dagger} \end{aligned} \tag{14.1}$$

where $\widehat{Y}^{T}$ is the transpose of matrix $\widehat{Y}$, and $\widehat{Y}^{\dagger}$ is the Moor-Penrose general invertible matrix [9, 10] of matrix $\widehat{Y}$. Integrating $A^{\dagger}$ and the estimated value $\hat{\boldsymbol{y}}$, $A^{\dagger}\hat{y}$ is obtained, which can be used to determine the estimated value $\hat{\boldsymbol{x}}$ of the original data instance. The above equation indicates that AIME is not applicable if $\widehat{Y}\widehat{Y}^{T}$ is not regular; however, the consideration that such cases rarely occur has been mentioned in a previous study [2].

14.3.2 Global Feature Importance in AIME

The approximate inverse operator $A^{\dagger}$ derived in Sect. 14.3.1 is a matrix of the features x number of classes that serves as the global feature importance. Considering the data analyzed in this study, which comprises two classes—stroke and non-stroke, the first column of $A^{\dagger}$ represents the behavior of the entire model in cases of non-stroke, hence, it helps recognize non-stroke instances when sorted by increasing absolute values. Similarly, if the second column is sorted by increasing absolute values, it helps recognize stroke instances. This trend is apparent in the overall behavior of the model. As demonstrated in a previous study [2], $A^{\dagger}$ acts as an approximate inverse operator of the black-box model, inferring $\hat{\boldsymbol{x}}$ from $\hat{\boldsymbol{y}}$. Given that $\hat{\boldsymbol{y}}$ is the confidence level of whether someone is having a stroke, it is reasonable to assume that $A^{\dagger}$ contains information regarding the contribution of features to determining stroke likelihood. This elucidates the roles of the first and second columns of $A^{\dagger}$ in identifying the occurrence of a stroke. It is worth noting that the output of AIME in these operations represents a feature contribution score that differs from the probabilistic estimates commonly found in other models.

14.3.3 Local Feature Importance in AIME

The approximate inverse operator $A^{\dagger}$ is derived by computing the $A^{\dagger}\hat{\boldsymbol{y}}$ term required to obtain the estimated $\hat{\boldsymbol{x}}$. Building upon this principle, the local feature importance in AIME is determined as follows [2]:

$$l = A^{\dagger}\hat{y} \circ x \tag{14.2}$$

where the length of the local feature importance vector l equals the number of features, with each value representing the feature importance and $\circ$ denoting the Hadamard product. The local feature importance vector $\boldsymbol{l}$ shows why $\hat{\boldsymbol{y}}$ is estimated given $\boldsymbol{x}$ and absolute values and how the importance coefficients can be extracted by sorting in descending order. This process facilitates the extraction of features from $\boldsymbol{x}$ that contribute to deriving $\hat{\boldsymbol{y}}$ when $\boldsymbol{x}$ is fed into the black-box model.

14.4 Experiments

14.4.1 Experimental Environments

The features of the stroke data [8] are presented in Table 14.1. The ID was omitted, whereas the gender, marital status, type of occupation, location of residence, smoking status, and stroke history were one-hot encoded. The one-hot encoded stroke instances constitute the objective variable matrix Y, whereas the non-encoded stroke instances form the explanatory variable matrix X. The dataset used in this study comprised information from 5,110 patients, with no missing data, which facilitated a more effective processing of categorical data. During the construction of the random forest model, 80% of the dataset served as training data, with the remaining 20% allocated for testing. This split ratio was chosen to accurately evaluate model performance. In this study, we used the random forest algorithm as the black-box model and implemented it using scikit-learn 1.2.2. Given the model-agnostic nature of AIME, explanations can be derived using various methodologies. The random forest model used in our study demonstrated outstanding performance, yielding an accuracy of 99.28%, precision of 98.58%, recall of 100%, F1 score of 99.28%, and ROC AUC score of 99.28%. The environment for this experiment was Google Colaboratory Pro+ in high-memory mode, using Python 3.10.12. Pre-AIME processing was implemented using NumPy 1.23.5, Pandas 1.5.3, and scikit-learn 1.2.2. Additionally, LIME [4] and SHAP [11], well-established XAI methods suitable for deriving local feature importance, were utilized. Finally, visualization tasks were accomplished using Seaborn 0.12.2 and Matplotlib 3.7.1.

Table 14.1 Feature set of the stroke prediction dataset

Features	Descriptions
Id	Unique identifier
Gender	Gender: "Male", "Female," or "Other"
Age	Age of the patient
Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
Ever_married	"No" or "Yes"
Work_type	"Children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
Avg_glucose_level	Average glucose level in blood
Bmi	Body mass index
Smoking_status	"Formerly smoked", "never smoked", "smokes" or "Unknown"*
Stroke	1 if the patient had a stroke and 0 if not

14.4.2 Experiment 1: Global Feature Importance in AIME

Figure 14.2 presents the outcomes of the global feature importance analysis conducted through AIME. While detailed results are not included due to space constraints, it is evident that unlike the random forest model, AIME distinctly identifies features that contribute to increased stroke risk as well as those that may offer protective effects, thereby offering a more nuanced perspective of the model's behavior.

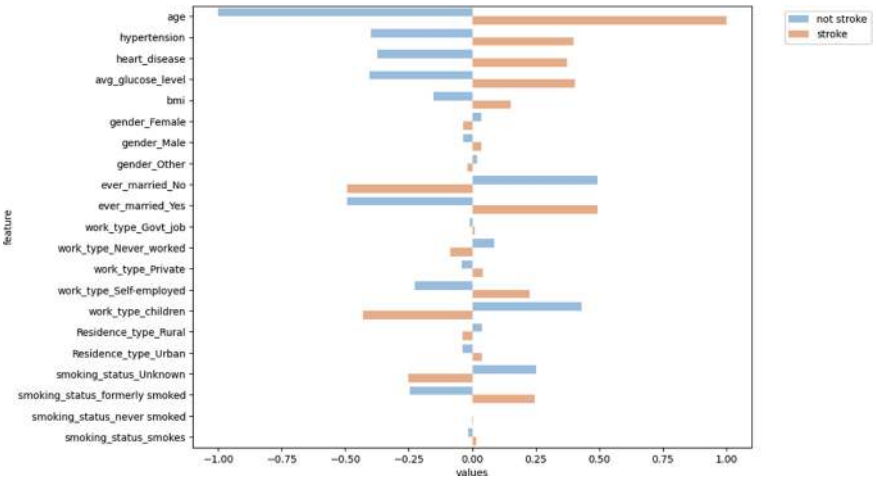


Fig. 14.2 Results of global feature importance for AIME

The results depicted in Fig. 14.2 highlight several factors that significantly influence the likelihood of stroke occurrence. These factors included the patients' age, history of hypertension or heart disease, average glucose levels, BMI, marital status, type of occupation, and smoking history. Specifically, an advanced age, suffering from hypertension or heart disease, an elevated average glucose level, higher BMI, self-employment, and a history of smoking were associated with an increased risk of stroke. Conversely, factors such as being unmarried, employment in occupations typically held by younger individuals (indicative of a younger age), and an unknown smoking status seemed to mitigate the risk.

The relationship between age and stroke has been extensively documented in numerous studies. For instance, Vokó et al. [12] reported hypertension as a significant risk factor for stroke among the elderly. Similarly, the correlation between heart disease and stroke has been subject to investigation. Bogousslavsky et al. [13] found that atrial fibrillation was implicated as the primary cause of stroke in only 18% of patients, while emboli originating from the heart were deemed the leading cause in 76% of cases. Average blood glucose levels have been associated with diabetes, a known precursor to stroke. Peng et al. [14] suggested that elevated and fluctuating mean glucose levels may serve as risk factors for stroke, even among individuals without a diabetes diagnosis. Furthermore, Kroll et al. [15] proposed a link between higher BMI, increased risk of ischemic stroke, and lower risk of hemorrhagic stroke. Although some studies have shown an inverse association, others suggest that BMI is a strong risk factor for total and ischemic stroke in women. Woodward et al. [16] identified a positive log-linear relationship between BMI and stroke risk. Regarding marital status, while it has been recognized that single individuals are generally considered to be at lower stroke risk compared to married individuals, the presence of a spouse may potentially elevate the risk. However, our findings, as illustrated in Fig. 14.2, suggest an opposite trend. For instance, Andersen et al. [17] found that marital status has minimal impact on stroke risk, whereas being divorced is linked to an increased risk, particularly among men. This implies that the 'ever married' category in our study includes individuals who have been divorced, and within it a group that is potentially at a higher risk. Additionally, Glymour et al. [18] highlighted the significant risk factor that a smoking spouse poses for stroke among nonsmokers and former smokers. Based on these findings, being 'ever married' in the present data cannot be medically related to stroke. Furthermore, several studies have suggested a correlation between self-employment and stroke risk. For instance, Krittanawong et al. [19] suggested that self-employment could be associated with elevated cardiovascular risk, including a higher incidence of stroke, particularly within the general population of the United States. Similarly, Colditz et al. [20] established smoking as a causative factor for stroke. This underlines the expectation of a more pronounced emphasis on smoking status in Fig. 14.2 concerning stroke risk. The analysis of these results indicates that, aside from marital status and smoking habits, the behavior of our model closely aligns with trends identified in previous medical research.

14.4.3 Experiment 2: Local Feature Importance in AIME

The data instance utilized in this experiment pertains to a 7-year-old female child who has never been married and is not employed. She has no history of hypertension or heart disease, an average glucose level of 88.6, and a BMI of 17.4. Her residence is located in an urban area, her smoking status is unknown, and her records indicate she has not experienced strokes. Figure 14.3 depicts the local feature importance results obtained with LIME [4], SHAP [11], and our method, AIME [2]. This analysis elucidates why a black-box model predicted no stroke for this data instance.

At first glance, the results from the LIME and SHAP analyses may suggest a rationale for classifying certain data points as indicative of stroke. However, instances where the value is recorded as '0' in the data instance influence this classification, posing a challenge for users attempting to comprehend the analysis, especially concerning factors like hypertension and heart disease. In contrast, the results derived from the AIME approach indicate that several demographic and health-related factors significantly influence stroke prediction. Specifically, factors such as age, average glucose levels, BMI, gender, marital status, type of occupation, and smoking history were identified as critical. Notably, being female, unmarried, employed in specific occupations, and having a history of smoking were associated with an increased risk of stroke. This analysis underscores the intricate interplay between various risk factors and their contribution to stroke likelihood. The importance of local features can be readily interpreted because an unmarried female with a high glucose level and high BMI is deemed to be at higher risk in the conventional medical studies discussed in Sect. 14.4.2. Figure 14.2 compares this local feature importance with global feature importance.

14.4.4 Discussion

The outcomes of Experiment 1 indicate that AIME's global feature importance underscores the significant influence of health-related factors such as age, blood glucose level, BMI, gender, marital status, occupation, and smoking history on stroke prediction. These findings align with existing medical research, except for marital status and smoking habits, where the model's consistency with established medical literature is observed.

The results of Experiment 2 also offer a straightforward explanation for the local feature importance outcomes of AIME and elucidate why the data were not identified as indicative of stroke. Conversely, findings from LIME and SHAP analyses may suggest reasons for classifying specific data points as strokes. However, instances labeled as "0" in Table 14.1 have been demonstrated to impact this classification, potentially complicating the interpretation of factors like hypertension or heart disease. This challenge may impede users' efforts to evaluate the effects of such factors. AIME addresses the shortcomings of LIME and SHAP.

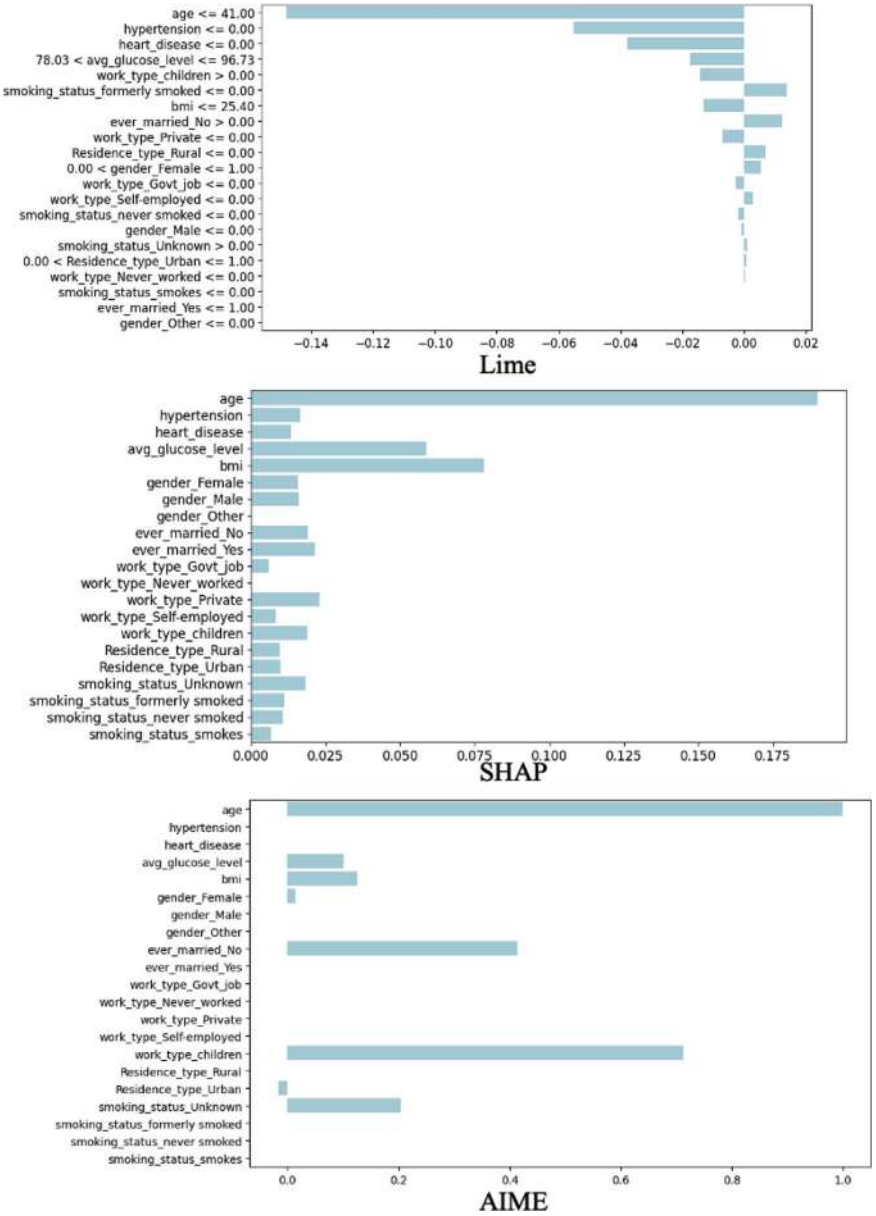


Fig. 14.3 Comparison of LIME, SHAP, and AIME (our method) results for local feature importance showing the explanation estimated in the black-box model as actual stroke in data that is stroke and correct

From these results, we could derive a clearer and simpler explanation of AIME, in contrast to the conventional LIME and SHAP methods. In the medical realm, where AI and machine learning are poised to become increasingly prevalent, the capacity to derive straightforward explanations like AIME holds promise for facilitating new discoveries and aiding decision-making in diagnoses. While this experiment utilized relatively simple data, future validation of AIME's explanation accuracy with more complex datasets is imperative.

14.5 Conclusion

In this study, we propose an explanation method for machine learning models utilizing XAI techniques, which has been the focus of our investigation thus far [2]. By integrating AIME with medical and stroke prediction data [8] and conducting experimental validation, we showcased the efficacy of our proposed methodology. Furthermore, we demonstrated that we can examine prior medical studies and assess whether the predictions rely on unrelated features by elucidating explanations for machine learning outcomes. We believe that this approach has the potential not only to enhance diagnostic accuracy for medical practitioners utilizing machine learning but also to serve as an effective tool for uncovering novel insights in medical research.

Despite its resilience against multicollinearity demonstrated in previous studies, AIME remains unable to explicitly elucidate the relationships between features. Therefore, our future endeavors will involve integrating and explicating these inter-feature relationships as part of the explanation process.

References

1. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019)
2. Nakanishi, T.: Approximate inverse model explanations (AIME): unveiling local and global insights in machine learning models. IEEE Access **11**, 101020–101044 (2023). https://doi.org/10.1109/ACCESS.2023.3314336
3. Islam, M., Hussain, I., Rahman, M., Park, S., Hossain, M.: Explainable artificial intelligence model for stroke prediction using EEG signals. Sensors (Basel, Switzerland) **22** (2022). https://doi.org/10.3390/s22249859.
4. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
5. ELI5 Documentation. https://eli5.readthedocs.io/en/latest/index.html. Last accessed 18 January 2024
6. Gandolfi, M., Boscolo, I., Gasparin, R., Cruciani, F., Vale, N., Picelli, A., Storti, S.F., Smania, N., Menegaz, G.: eXplainable AI allows predicting upper limb rehabilitation outcomes in sub-acute stroke patients. IEEE J. Biomed. Health Inf. (2022). https://doi.org/10.1109/JBHI.2022.3220179

7. Bhattacharya, G.T., Maddikunta, P., Hakak, S., Khan, W., Bashir, A., Jolfaei, A., Tariq, U.: Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. Multimed. Tools Appl. **81**, 41429–41453 (2020)
8. Hassan, A.: Stroke Prediction Dataset. IEEE Dataport (2023). https://doi.org/10.21227/mxfb-sc71. Last accessed on 18 January 2024
9. Moore, E.H.: On the reciprocal of the general algebraic matrix. Bull. Am. Math. Soc. **26**, 294–300 (1920)
10. Penrose, R.: A generalized inverse for matrices. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 51, no. 3, pp. 406–413. Cambridge University Press, Cambridge (1955). https://doi.org/10.1017/S0305004100030401
11. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30** (2017)
12. Vokó, Z., Bots, M., Hofman, A., Koudstaal, P., Witteman, J., Breteler, M.: J-shaped relation between blood pressure and stroke in treated hypertensives. Hypertension **34**(6), 1181–1185 (1999). https://doi.org/10.1161/01.HYP.34.6.1181
13. Bogousslavsky, J., Melle, G., Regli, F., Kappenberger, L.: Pathogenesis of anterior circulation stroke in patients with nonvalvular atrial fibrillation. Neurology **40**, 1046–1046 (1990). https://doi.org/10.1212/WNL.40.7.1046
14. Peng, X., Ge, J., Wang, C., Sun, H., Ma, Q., Xu, Y., Ma, Y.: Longitudinal average glucose levels and variance and risk of stroke: a Chinese cohort study. Int. J. Hypertens. (2020). https://doi.org/10.1155/2020/8953058
15. Kroll, M., Green, J., Beral, V., Sudlow, C., Brown, A., Kirichek, O., Price, A., Yang, T., Reeves, G.: Adiposity and ischemic and hemorrhagic stroke. Neurology **87**, 1473–1481 (2016). https://doi.org/10.1212/WNL.0000000000003171
16. Woodward, M., Fang, X., Gu, D., et al.: Impact of cigarette smoking on the relationship between body mass index and coronary heart disease: a pooled analysis of 3264 stroke and 2706 CHD events in 378579 individuals in the Asia Pacific region. BMC Public Health **9**, 294 (2009). https://doi.org/10.1186/1471-2458-9-294
17. Andersen, K., Olsen, T.: Married, unmarried, divorced, and widowed and the risk of stroke. Acta Neurol. Scand. **138**, 41–46 (2018). https://doi.org/10.1111/ane.12914
18. Glymour, M., DeFries, T., Kawachi, I., Avendano, M.: Spousal smoking and incidence of first stroke: the health and retirement study. Am. J. Prev. Med. **35**(3), 245–248 (2008). https://doi.org/10.1016/j.amepre.2008.05.024
19. Krittanawong, C., Kumar, A., Wang, Z., Baber, U., Bhatt, D.: Self-employment and cardiovascular risk in the US general population. Int. J. Cardiol. Hypertens. **6** (2020). https://doi.org/10.1016/j.ijchy.2020.100035
20. Colditz, G., Bonita, R., Stampfer, M., Willett, W., Rosner, B., Speizer, F., Hennekens, C.: Cigarette smoking and risk of stroke in middle-aged women. N. Engl. J. Med. **318**(15), 937–941 (1988). https://doi.org/10.1056/NEJM198804143181501

Chapter 15
Improving Membership Inference Attacks Against Classification Models

Shlomit Shachor, Natalia Razinkov, Abigail Goldsteen, and Ariel Farkash

Abstract Artificial intelligence systems are prevalent in everyday life, with use cases in retail, manufacturing, health, and many other fields. With the rise in AI adoption, associated risks have been identified, including privacy risks to the people whose data was used to train models. Assessing the privacy risks of machine learning models is crucial to making knowledgeable decisions on whether to use, deploy, or share a model. A common approach to privacy risk assessment is to run one or more attacks against the model and measure their success rate. We present a novel framework for improving the accuracy of membership inference attacks against classification models. Our framework takes advantage of the ensemble method, generating many specialized attack models for different subsets of the data. We show that this approach achieves better performance than either a single attack model or an attack model per class label, on both classical and language classification tasks.

This work was performed as part of the NEMECYS project, which is co-funded by the European Union under grant agreement ID 101094323, by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee grant numbers 10065802, 10050933 and 10061304, and by the Swiss State Secretariat for Education, Research and Innovation (SERI).

S. Shachor · N. Razinkov · A. Goldsteen (✉) · A. Farkash
Data Security and Privacy, IBM Research, Haifa, Israel
e-mail: abigailt@il.ibm.com

S. Shachor
e-mail: shlomiti@il.ibm.com

N. Razinkov
e-mail: natali@il.ibm.com

A. Farkash
e-mail: arielf@il.ibm.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_15

15.1 Introduction

Artificial intelligence (AI) systems have become prevalent in everyday life. AI is used in retail, security, manufacturing, health, finance, and many more sectors to improve or even replace existing processes. However, with the rise in AI adoption, different risks associated with AI have been identified, including privacy risks to the people whose data was used to train the models. In addition to fundamental societal harm, these risks can result in negative brand reputation, lawsuits, and fines. This has given rise to the notion of Trustworthy or Responsible AI.

A key aspect of Responsible AI is the ability to assess (and later mitigate) these risks. Assessing the privacy risk of machine learning (ML) models is crucial to enable well-informed decision-making about whether to use a model in production, share it with third parties, or deploy it in customers' homes. The most prevalent approach to privacy risk assessment is to run one or more known attacks against the model and measure how successful they are in leaking personal information.

The most common attack used in model assessment is called *membership inference*. Membership inference attacks (MIA) aim to violate the privacy of individuals whose data was used in training an ML model by attempting to distinguish between samples that were part of a target model's training data (called members) and samples that were not (non-members), based on the model's outputs. These can be class probabilities or logits (for classification models), the model's loss, or activations from internal layers of the model (in white-box attacks). Most attacks choose one or more of these features and train a binary classifier to try to distinguish between members and non-members. The success of such attacks can be measured using standard ML metrics such as Accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), or as suggested recently by Carlini et al. [2], by the True Positive Rate (TPR) at low False Positive Rate (FPR).

In this paper, we present a novel framework for MIA against classification models that takes advantage of the ensemble method to generate many specialized attack models for different subsets of the data. This ensemble method can be applied to any existing model-based attack, improving its results by up to 14% (according to our experiments) when compared to a single attack model or an attack model per class label, both on classical and language classification tasks. This improvement stems from the specialization of each attack model to the specific data spilt it was trained on, based on a grid search of the best combination of attack model architecture, input features to the attack, and scaling method. This results in each model being best suited to identify membership leakage for a specific subset of data. We evaluated our method both on language models that have an explicit classification head and generative models that can respond to classification prompts or instructions (such as the flan-UL2 model).

Our method can cater to both privacy audit mode, in which an organization assesses the privacy vulnerability of their own models, and attack mode, where the real training data is unknown to the attacker. For the latter, a preceding step of generating shadow models and data is required [15].

In the realm of large language models (LLM), membership inference can be assessed for different phases of the model's development, namely the pre-training and fine-tuning stages. Pre-training is largely performed on publicly available datasets, and the data used to train a model is often also public knowledge. Fine-tuning is typically performed on a smaller, proprietary dataset. It is therefore more common to look at the fine-tuning phase in the context of MIAs. However, this framework can be applied to either of these phases.

The paper starts by surveying relevant prior work in Sect. 15.2. Next, we describe our framework for improved membership inference attacks based on small specialized attack models in Sect. 15.3. We present our evaluation results in Sect. 15.4. We discuss those results in Sect. 15.5 and conclude in Sect. 15.6.

15.2 Related Work

There are several types of privacy (inference) attacks against ML models, including membership inference, attribute inference, model inversion, database reconstruction, and most recently, training data extraction from generative models. The most commonly researched and employed attack is the membership inference attack, with dozens of papers published each year [5], and implementations being made available in open-source privacy assessment frameworks [7, 12].

MIAs attempt to distinguish between members, which were part of a target model's training data, and non-members. MIAs have been extensively studied in the context of classification models and in the black-box setting, where the model internals are unknown to the attacker. The first MIAs were either threshold-based [17] or employed binary classifiers trained to distinguish between members and non-members based on model outputs [15]. For example, these outputs may include class probabilities or logits (for classification models), the model's loss, and possibly also activations from internal layers of the model (in white-box attacks) [11]. To generate labeled (member/non-member) data to train the attack classifier, without knowledge of the true member samples of the attacked model, shadow models are commonly used [15].

In the past few years, investigations have begun into MIA in the context of large language models (LLM), starting with embedding models and masked language models [8, 10, 16]. Shejwalkar et al. [14] looked at a similar setting as ours, focusing on NLP classification models. They proposed mostly threshold-based attacks, examining different features that can be used to distinguish between members and non-members. Jagannatha and Rawat [6] focused specifically on language models that were fine-tuned for the medical domain, including classification tasks such as MedNLI, employing both black-box and white-box attacks. Their black-box attack applied thresholds to the training error of samples. More recently, Likelihood Ratio Attacks (LiRA) have been proposed [2], which compare target model scores to those obtained from a reference model trained on similar data. Mattern et al. [9] tried to relieve the assumption that an adversary has access to samples closely resembling the original training data by utilizing synthetically generated neighbor texts.

Some works on MIA in the language domain differentiate between sample-level MIA, which treats each text sequence/document in the training data separately, and user-level MIA, which groups together samples originating from the same person or source. In this work, we focus solely on the sample level, which is usually considered a harder problem.

Most existing approaches to MIA that employ a classification model use either a single attack model for the entire dataset or a separate attack model per class. Ensembles have been used in a few cases in the context of adversarial (evasion) attacks to generate more robust adversarial examples [1, 3]. Salem et al. [13] used multiple shadow models to compensate for a lack of knowledge of the target model algorithm; however, these multiple models were only used to generate multiple shadow datasets, which were then combined to train the attack model.

15.3 The Framework

We propose a method for improving the performance of model-based membership inference attacks by splitting the initial member and non-member datasets into multiple small, non-overlapping subsets, used to train different attack models. Thus, multiple specialized attack models are generated for small pieces of the data, where each model is best in identifying membership leakage for that piece. To find the best possible attack model for each subset, many different combinations of model type, scaling, and input features are tried. The best combination is selected based on the highest score for the specific metric being measured, e.g., Accuracy, AUC-ROC, or TPR@low FPR. Aggregating the results from those multiple attack models can better reveal the real leakage of the target model.

The source of the member and non-member data input to this process is irrelevant and can come from either shadow datasets (in attack mode) or from the actual training and test sets of the target model (in audit mode).

Our method can be used for any model that can perform classification tasks. This includes classical ML models, such as a decision tree or random forest, language classification models, and even generative models that were fine-tuned for text classification tasks.

15.3.1 Use of Small Specialized Attacks

The high-level flow of the proposed framework is depicted in Fig. 15.1. The first step is to split the member and non-member datasets into non-overlapping subsets and randomly assign member/non-member pairs from those subsets (the pairs remain constant). In subsequent phases, each pair serves to generate a specialized attack model.

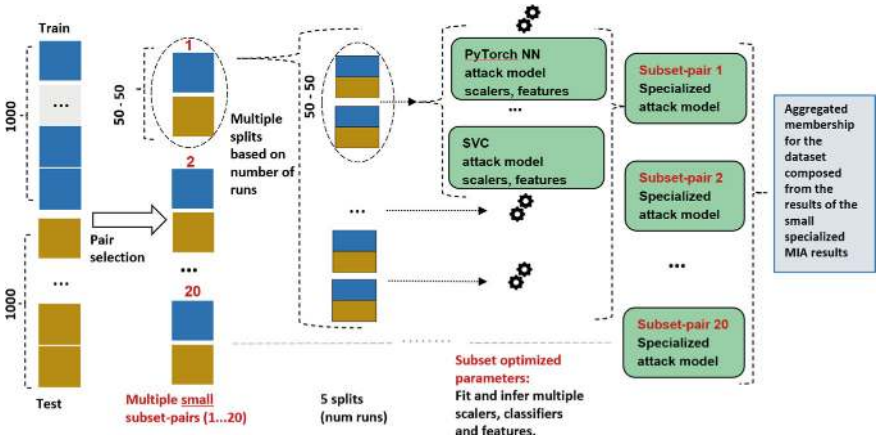


Fig. 15.1 High-level overview of the framework for small specialized MIA models

Each pair is then split into two halves, one for fitting the attack model and the second for inferring membership. This splitting process is done multiple times for each pair. Our experiments show that even for the same pair of member and non-member subsets, the half used for training the attack model has significant impact on the model's ability to infer membership. Thus, we perform this split multiple times, eventually using the attack that achieves the best result (highest leakage). This allows us to generate attack models that are even more specialized and powerful.

For each split, the data is passed through the target model and various features are calculated based on the model outputs. The features may include predicted labels, losses, class probabilities or class-scaled probabilities, entropy, modified entropy [14], scaled class logits [2], etc. The list of features to use in each attack may be pre-configured or optimized based on the best attack performance for each pair.

The features are then scaled using different types of scalers (e.g., robust, min-max) and used to train multiple types of binary classifiers (e.g., random forest, k-nearest neighbors, decision tree, etc.) with possibly different combinations of input features. The specific combination of scaler, classifier, and features is selected to achieve the best attack performance (worst-case privacy leakage) for each pair. This can be, for example, the best accuracy or the best AUC-ROC score. This process yields different attack models (based on different combinations) for each subset. The results of all of these selected attacks are averaged, so that it fairly represents the performance on all of the data assessed.

In addition, this whole flow can be performed multiple times (called instances) on different random samples of the entire member and non-member datasets. The sample size can be substantially smaller than the entire dataset size, which is especially beneficial when the datasets are very large. This improves the performance of the overall assessment process, while providing good coverage of the data. For each instance of the flow that is run, the results of all subset pairs are averaged, and finally

the results of all instances are aggregated. This aggregation can be done in different ways. In the evaluation section, we show aggregation based on average attack results.

Even though our main goal is to perform privacy evaluation of models (audit mode), the framework can also be used in attack mode where the true membership status of samples is unknown when performing inference. In this case, typical ensemble aggregation methods, such as majority voting, can be used.

The parameters controlling the process such as the subset size, number of subsets, number of runs per subset, number of instances, and type of aggregation can be easily configured.

15.4 Evaluation

Since we are targeting the privacy audit scenario and want to simplify the evaluation, we use the known training and test data of the model when conducting our experiments. All results presented in this section were performed in the same manner to enable a fair comparison.

We experimented with several LLM architectures and datasets: two classification models from huggingface—textattack/bert-base-uncased-SST-2, fine-tuned on glue-SST2[1] (denoted BS), and textattack/roberta-base-CoLA, fine-tuned on glue-CoLA[2] (RC); one generative model - google/flan-ul2 trained with glue-CoLA (FC) and glue-SST2 (FS); and a roberta-base model fined-tuned with the rotten tomatoes dataset[3] (RR). All of these models were fully fine-tuned on the given dataset. We also used in our evaluation a roberta-base model fined-tuned on rotten tomatoes using parameter efficient fine-tuning with LoRA [4] (RR-L). Finally, to assess the effectiveness of our method on models with a privacy defense, we also evaluated the roberta-base model after applying differentially private fine-tuning using DP-LoRA [18] with $\epsilon = 2$ (RR-DP). Table 15.1 presents the dataset sizes and the accuracy of all evaluated models for the test and training data. For the SST2 and CoLA datasets, we used the validation set as test data for our evaluation, due to a lack of true labels in the test datasets.

In our experiments, we set the subset size to 50, the number of runs to 5, the number of instances to 50, and the size of the member and non-member samples for each instance to 1000 each (872 for SST-2).

We compare the ensemble method both to training a single attack model on the entire dataset and class-based attacks, where a separate attack model is trained per class label. The single-model or model-per-class attacks serve as our baseline. For these baseline attacks we employed the exact same attack implementation but without the stage of dividing the data received as input into separate non-overlapping subsets. We also used 5 runs and 50 instances per experiment in each of these attacks.

[1] https://huggingface.co/datasets/glue/viewer/sst2.

[2] https://huggingface.co/datasets/glue/viewer/cola.

[3] https://huggingface.co/datasets/rotten_tomatoes.

Table 15.1 Accuracy of the evaluated models on train and test data

Model	Train accuracy	Test accuracy	Train set size	Test set size
RC	0.948	0.850	8551	1043
BS	0.986	0.924	67348	872
FC	0.930	0.864	8551	1043
FS	0.960	0.964	67348	872
RR	1.000	0.877	7500	1066
RR-L	0.978	0.889	7500	1066
RR-DP	0.859	0.849	7500	1066

For each model and its corresponding data, we conducted six experiments: using a single attack model for the entire dataset (S01) and a single model for class 0 (S0) and for class 1 (S1); and using many small specialized attack models for the entire dataset (M01), for class 0 (M0), and for class 1 (M1). In all of these experiments, we used the following features to train the attack models: true labels, predicted labels, class-scaled probabilities, class-scaled logits, losses, and modified entropy [14].

For google/flan-UL2, a generative model, we used commonly available prompts for CoLA and SST2, requesting the model to classify the linguistic correctness of a sentence or its sentiment, respectively. For CoLA we used: "Sentence: {sentence}. Would a linguist rate this sentence to be acceptable linguistically? Options: acceptable, unacceptable. Answer:", and for SST2: "Sentence: {sentence}. What is the sentiment of this sentence? Options: positive, negative. Answer:".

We instructed the model to generate a score structure instead of just text, and used low temperature mode to ensure determinism. The score structure was used to calculate the features mentioned above (e.g., probabilities and entropy). In addition, we calculated the perplexity for each of the choices ("positive" and "negative" for SST2; "acceptable" and "unacceptable" for CoLA) and used it as an additional feature.

For the different attack model architectures, we employed the following model types from scikit-learn[4]: RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, DecisionTreeClassifier, and KNeighborsClassifier, all with the default parameters, as well as SVC (C-Support Vector Classification) with rbf, sigmoid, and poly kernels. In addition we employed the XGBClassifier from the xgboost package,[5] and a PyTorch[6] Neural Network (NN) with three fully connected layers of sizes 512, 100 and 64 respectively, and a sigmoid activation. It was trained for 100 epochs with a batch size of 100, using an Adam optimizer with initial learning rate of 0.0001.

For scaling the input features to the attack, we varied between the scikit-learn scalers: StandardScaler, MinMaxScaler, and RobustScaler. As mentioned earlier,

[4] https://scikit-learn.org/stable/.

[5] https://github.com/dmlc/xgboost.

[6] https://pytorch.org/.

the best combination of model, scaler and input features are selected in each run, and the results aggregated to yield the best overall attack score.

15.4.1 Results

The average TPR@low FPR (1%), AUC-ROC, and Accuracy scores across all instances are presented in Fig. 15.2 (in percentages). We use a FPR of 1% and not 0.1% because our datasets are not so large and it was not always possible to achieve lower FPR values with these smaller datasets. It is clear that in all cases, the many small attacks method (green) outperforms the single attack (blue). Detailed scores are brought in Appendix 15.7 (Table 15.2).

Our experiments show improvements of between 4–11% in average accuracy and AUC-ROC and up to 1.76% in TPR@low FPR for the undefended models, across all datasets and attacks tested, even when compared with class-based methods. Surprisingly, for the defended model (RR-DP), even though previous attacks seem mostly mitigated with accuracy and AUC-ROC scores ranging from 53% to 55% (which is very close to random guessing), our attack is able to achieve a significant advantage of up to 14% above the baseline attacks. In this case, our attack also achieved the highest improvement in TPR@low FPR of 1.79% above the baseline. This shows that our method is especially advantageous against models to which a privacy defense has been applied. Nonetheless, the accuracy and AUC-ROC scores achieved by our attack on the defended model were still lower than for the undefended ones.

15.5 Discussion

In this framework and its evaluation, several design choices were made based on logical or performance reasons. For example, the pairs of member and non-member subsets were assigned randomly and fixed throughout the experiment (within a single instance). Testing all possible combinations of pair assignments to find the best match would likely increase the attack's success rate even more. Another possibility is analyzing the dataset to try to find characteristics that can be leveraged when splitting the data and assigning pairs.

Moreover, when combining the results of the subsets, we always used averaging to enable a fair comparison between the attacks run on the entire dataset and the small attacks. However, it is also possible to choose the best subset.

Varying the input features to the attack did not have a significant effect on the success rate. Rather, the main advantage stems from the use of many different attacks for different data subsets and the specialization of the attack models. It is worth noting that in most cases, either the SVC or the NN model were the ones to achieve the best attack performance.

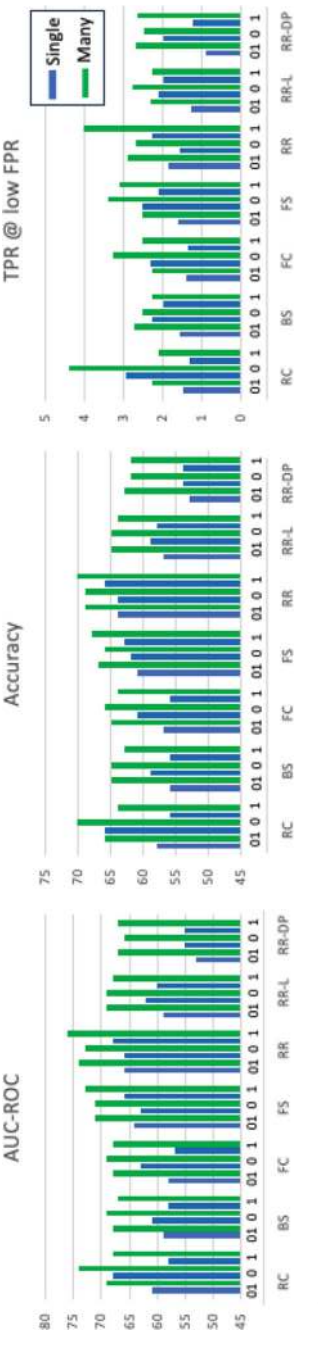


Fig. 15.2 Comparative results between single attack and multiple small attacks across models and datasets. Blue lines represent single attack, green lines represent many attacks. Each pair of adjacent lines represent the same experiment: both classes together (01), and per class (0 or 1 respectively)

Table 15.2 Average performance metrics across all instances for single vs. many attack models (TPR@low FPR (1%)|AUC-ROC|Accuracy), all in percentages.

Model	S01	**M01**	S0	**M0**	S1	**M1**
RC	1.49|61|58	**2.26|69|66**	2.93|68|66	**4.39|74|70**	1.32|58|56	**2.10|68|64**
BS	1.58|59|56	**2.75|68|65**	2.28|61|59	**2.53|69|65**	1.97|58|56	**2.26|67|63**
FC	1.41|58|57	**2.26|68|65**	2.34|63|61	**3.29|69|66**	1.35|57|56	**2.51|68|64**
FS	1.60|64|61	**2.51|71|67**	2.53|63|62	**3.41|71|66**	2.10|66|63	**3.11|73|68**
RR	1.86|66|64	**2.92|74|69**	1.57|66|64	**2.68|73|69**	2.27|68|66	**4.03|76|70**
RR-L	1.28|59|57	**2.30|69|65**	2.10|62|59	**2.79|69|65**	1.98|60|58	**2.26|68|64**
RR-DP	0.91|53|53	**2.70|67|63**	1.98|55|54	**2.49|66|62**	1.26|55|54	**2.66|67|62**

15.6 Conclusion and Future Work

We presented a novel method for running membership inference attacks that divides the data into small subsets and trains specialized attack models for each subset. This method significantly improves the success rate of attack models trained on the entire data or per class label. It can be applied to both classical models as well as large language models that perform classification tasks and even succeeds in attacking models defended using differential privacy.

During our experimentation, we saw indications that the prompt used when assessing generative models has a significant effect on the success rate of the attack. We plan to investigate this further and perhaps add it as an additional source of variability in the framework.

Moreover, we plan to check the viability of this approach for other types of privacy attacks such as attribute inference and other types of target models besides classification.

15.7 Average Attack Results Across Instances

Table 15.2 presents the average attack scores (TPR@low FPR (1%), AUC-ROC, and Accuracy) of all the instances in our experiments.

References

1. Cai, Z., Tan, Y., Asif, M.S.: Ensemble-based blackbox attacks on dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4045–4055 (2023)
2. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 1897–1914. IEEE (2022)

3. Fu, Z., Cui, X.: Elaa: an ensemble-learning-based adversarial attack targeting image-classification model. Entropy **25**(2), 215 (2023)
4. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=nZeVKeeFYf9
5. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: a survey. ACM Comput. Surv. 54(11s) (Sept 2022). https://doi.org/10.1145/3523273
6. Jagannatha, A., Rawat, B.P.S., Yu, H.: Membership inference attack susceptibility of clinical language models (2021). arXiv:2104.08305
7. Kumar, S., Shokri, R.: Ml privacy meter: aiding regulatory compliance by quantifying the privacy risks of machine learning. In: Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs) (2020)
8. Mahloujifar, S., Inan, H.A., Chase, M., Ghosh, E., Hasegawa, M.: Membership inference on word embedding and beyond (2021). arXiv:2106.11384
9. Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., Berg-Kirkpatrick, T.: Membership inference attacks against language models via neighbourhood comparison (2023). arXiv:2305.18462
10. Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., Shokri, R.: Quantifying privacy risks of masked language models using membership inference attacks. arXiv preprint arXiv:2203.03929 (2022)
11. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 739–753. IEEE (2019)
12. Nicolae, M.I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., Edwards, B.: Adversarial robustness toolbox v1.2.0. CoRR 1807.01069 (2018). https://arxiv.org/pdf/1807.01069
13. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models (2018). arXiv:1806.01246
14. Shejwalkar, V., Inan, H.A., Houmansadr, A., Sim, R.: Membership inference attacks against nlp classification models. In: NeurIPS 2021 Workshop Privacy in Machine Learning (2021)
15. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE Computer Society, Los Alamitos, CA, USA (May 2017). https://doi.ieeecomputersociety.org/10.1109/SP.2017.41
16. Song, C., Raghunathan, A.: Information leakage in embedding models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 377–390 (2020)
17. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pp. 268–282 (2018)
18. Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H.A., Kamath, G., Kulkarni, J., Lee, Y.T., Manoel, A., Wutschitz, L., Yekhanin, S., Zhang, H.: Differentially private fine-tuning of language models. In: ICLR (2022)

Chapter 16
AI, Law and (Neuro-) Rights as New Human Rights?

Melania D'Angelosante

Abstract This paper tries to provide some answers to the question whether the establishment of the so-called 'neuro-rights' is necessary for the protection of people from possible risks related to the use of artificial intelligence (AI). An evident relationship exists between the development of new technologies and law, being law instrumental to the regulation of life in human communities, for their own survival. On the one hand, new technologies affect traditional legal categories. On the other hand, there is a frequent need to carry out a technological impact analysis of the laws in force or to be enacted, since they regulate events conditioned by the application of technologies, and due to the need for a preliminary technological education of decision-makers, as to allow the proper knowledge of the facts to be regulated. The paper will be divided into a first part dedicated to the description of the context, a second part dedicated to highlighting the main problematic issues at present arising from the relationship between the role of law and the development of new technologies, and a third part where partial conclusions will be drawn by developing some general reflections.

16.1 Introduction

The word 'neuro-rights' was for the first time used in 2015 by Marcello Ienca, in order to indicate the need to establish a kind of habeas mentem, aimed at safeguarding the proper use of neural data (after the habeas corpus, aimed at safeguarding the body of the persons from coercive acts, and after the habeas data, aimed at safeguarding the autonomy of information in the information society) [1]. In that context, the following rights were identified: (1) the right to cognitive freedom (i.e. to decide autonomously and freely on the 'disclosure' and/or enhancement of the human beings' cognitive processes), (2) Right to mental privacy (i.e. to preserve mental information, other than neural information, from unauthorised access), (3) Right to mental integrity (i.e. to protect mental activity from unauthorised and harmful manipulation), (4) Right

M. D'Angelosante (✉)
University G. d'Annunzio of Chieti-Pescara, ITA, Viale Pindaro 42, 65121 Pescara, Italy
e-mail: melania.dangelosante@unich.it

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_16

to psychological continuity (i.e. to preserve personal identity based on brain activity and the continuity of mental life from unauthorised external alterations) [2].

After that, in 2017, the Columbia University research group coordinated by the neuroscientist Prof. R. Yuste specified the contents of the so-called neuro-rights, identifying: (1) the right to mental privacy and consent; (2) the right to personal freedom/ personal identity (and psychological continuity); (3) The right to mental integrity and freedom of choice; (4) The right to equal access to mental enhancement; (5) The right to protection from algorithmic errors.

The development of new technologies, together with increasing their essentiality for human activities, also influenced the methods of studying the relationship between these technologies and law: for instance, in the Anglo-American area, a model emerged where the analysis of norms is related to (and/or conditioned by) a preliminary or contextual analysis of the behaviour of users in their interaction with technological devices whose use requires (or poses problems of) consent to the processing of personal data.

On the one hand, new technologies affect traditional legal categories.

On the other hand, there is a frequent need to carry out a technological impact analysis of existing norms or norms to be enacted, when they regulate facts conditioned by the application of technologies [3]: as a consequence, a need for a preliminary technological education of public decision-makers emerges, as a tool for implementing their understanding and comprehension of the facts to be regulated [4].

This paper tries to propose some answers to the question whether the establishment of the so-called 'neuro-rights' is necessary for the protection of people from possible threats related to the use of artificial intelligence (AI). An evident relationship exists between the development of new technologies and law, being law instrumental to the regulation of life in human communities, for their own survival.

As a consequence, usually law 'comes after' the phenomena of associated life: for example, the economy, the development of hard sciences, and the spread and evolution of new technologies. Thus, as a rule, the law comes later, with the aim of disciplining these phenomena, which, in turn, and on the contrary, originate spontaneously from the needs that communities of people express.

In this perspective, the present paper will consist of a first part dedicated to the description of the context, a second part dedicated to highlighting the main problematic issues at present arising from the relationship between the role of law and the development of new technologies, and a third part where partial conclusions will be drawn by developing some general reflections.

16.2 Preliminaries: Neurotechnology and Law

The most recent developments in studies on the use of technologies by law concern neurolaw research, focussed on the possible limits of such use, and in particular on the possible use of neurosciences, especially the latest medical technologies, by law: think to the relationship between brain damage/pathology—revealed by the most

advanced brain images—and the investigation of human behaviour. In particular, neurosciences consist of a set of scientific disciplines all focussed, from different perspectives, to the study of the nervous system. In turn, neurolaw studies deal, from an interdisciplinary approach, with the legal implications of the evolution of neurosciences and the consequent use/application of the substances/machines/technologies produced in the course of this evolution.

In this scenario, the use of personal biometric data, to which behavioural data can also be traced back, is frequent: think to the detection of eye movement by the devices present in some vehicles for the monitoring of the driver's level of attention and the implementation of the consequent actions for road safety. In fact, biometric techniques make it possible to recognise the identity of people by means of the evaluation of physiological or behavioural characteristics that are unique and exclusive to each individual. The existence and life cycle of biometric techniques depend on technological development: in terms of spread, evolution, and obsolescence [5].

Moreover, the source of brain manipulation by the most innovative products of science and technology may be heterogeneous, and not only AI-centred: for instance, some neuro-drugs improve attention, enhance cognitive faculties, alter sleep–wake rhythms and thus also affect thought processing. In the medical field, however, a consolidated system of ex ante, in itinere, and ex post control of such treatments exist, in addition to the general regime of informed consent. At present, pharmacological 'brain-doping' is prohibited, except on clinical grounds, i.e. for the treatment of pathologies [6]. However, 'brain-doping' (neuro-enhancement) also extends beyond the pharmacological sphere, including, in a broader meaning, the use of substances or technologies that, by acting on the central nervous system, make it possible to improve mental performance, for example by reducing the brain's recovery time based on sleeping, thus allowing waking sessions for longer than average, or by improving the ability to cope with fatigue/tiredness, or by increasing the body's precision and speed of responding to brain impulses. The substances suitable for inducing such effects are, therefore, not only drugs, but also some active ingredients from the food or homeopathic/herbal industry (think caffeine). Among the technologies, however, one can include the above mentioned transcranial continuous current stimulation, which is at present allowed and sometimes used in competitive sports.

However, no correspondence exists between the actual control systems in the medical field and the monitoring of the 'treatments' to which people are often unknowingly subjected by interacting with AI systems, which can pick up signals and send impulses, establishing communications.

Think of the so-called wellness technologies, such as those consisting of hi-tech wearables, which, as a rule, are directly usable by the consumer regardless of preventive control by the competent authorities: technologies capable of measuring heartbeat, blood oxygenation level, physical activity, and energy expenditure, and electrical activity of the brain belong to this category. Furthermore, some functions of smartwatches or applications for smartphones, or devices capable of performing the above mentioned transcranial direct current stimulation (usually consisting of wearable solutions similar to earphones), belong too to the same category.

It is certainly necessary to keep the biological study of the brain and the use of science and neurotechnologies for therapeutic purposes (on the one hand) distinct from their use for the investigation and conditioning of cognitive activities (on the other). That is, to distinguish their therapeutic use from their use for cognitive investigation and/or cognitive enhancement. But this second frontier generates questions of ethical sustainability, even before that of regulation by law.

However, as there are strong implications between neurological activity, consciousness, and identity, the study of the brain cannot be conducted just on the basis of biological parameters [7].

16.3 Discussion. Dystopian Scenarios: Do We Need Neuro-Rights?

The current debate not only focuses on the need of the so-called 'neuro-rights', with the aim of avoiding/limiting the threats to which some AI applications could expose fundamental rights.

In fact, the increasing functionality of technologies in human activities also implies questions of techno-ethics, based on the search for limits to technological development, defining the opposite scenarios of the upward assimilation of machines to man, and downward assimilation of man to machines [4].

In particular, the impact of 'machines' on the cognitive and evaluating faculties of their users is under discussion. The main reason of this phenomenon is that one of the peculiar aspects of AI consists in the ability of software to 'learn' from experience, so as to investigate and discover, through tracking activities, the preferences of its user (or consumer/voter), whom profiling allows the orientation of his future choices (so-called social engineering) [4], but also the driving of the preferences always in the same direction/perimeter (so-called cluster-effect) [8].

Also, the matter of the processing of unconscious personal data exists, since this information is 'snatched' by the AI-system in the absence of consciousness of the individuals to which it refers about its 'capture', its use, and the consequences of this use [9].

The most recent studies on the application of new technologies by law concern neurolaw research, which focuses on the limits of such use, on the use of neurosciences and medical technologies in the judicial field: think to the relationship between a pathology revealed by the most advanced neuroimaging and the investigation of liability [4].

One of the most important questions regards the legal consequences of the possible capability of AI to discover unexpressed mental conditions, to discover the way by virtue of which information is acquired by the brain and the will is produced (so-called brain-reading), and finally to manipulate this process.

The neurotechnologies that can be used for this purpose are the neuroimaging-technologies (aimed at the representation of brain activity), the human-machine interfaces (which allow impulses to be communicated to machines, for instance via the electrical activity detectable from the scalp), and the brain-machine interfaces (which detect brain signals by placing them in communication with external devices).

As we have seen, in the medical field treatments affecting brain activity are subject to preventive control and to monitoring of their effects; moreover, they are only permitted for specific clinical needs. On the other hand, wellness technologies are, as a rule, directly usable by the consumer: think of wearables capable of measuring brain activity, in order to transmit impulses that can foster concentration or relaxation.

According to some scholars, the so-called neuro-rights would safeguard the individuals' self-determination (mental integrity) and freedom of information/evaluation, especially in the face of possible threats related to AI [9, 10].

Some possible threats consist of a prospective overlapping between human and algorithmic action, making more difficult the distinction between the human willing process and its conditioning by the device [11].

A preliminary issue is the preservation of personal identity, the essential core of which is considered to be the mind as an inaccessible space. This is the basis, for instance, of the impossibility of using technologies such as lie detectors in criminal proceedings, in order to avoid the violation of moral freedom, which is essential for the protection of human dignity [12].

But new technologies also contributed to the fragmentation and dynamization of the concept of identity: beyond the personal identity, one could at present find also the narrative identity (defined by the paths traced by search engines), transactive identity (resulting from profiling), and predictive identity, which are all capable to change very fast (compared to the personal identity). The latter (predictive identity) is strictly related to the protection of mental privacy, i.e. the right to prevent unexpressed personal information from being captured through disclosures of the mental condition, and thus processed in a predictive way.

The thesis according to which the so-called neuro-rights are not necessary is based on the will to prevent an inflation of guarantees undermining the existing degree of protection [12].

The scenario of the already existing sources of law which can be interpretated in an extensive way as to include also the protection of the so-called neuro-rights is broad. It is possible to mention, for the Italian legal system, Articles 2 and 21 of the Italian Constitution [12], 188 et seq. and 64 p. 2 of the Code of Criminal Procedure, the rules on contractual invalidity, on consumer protection. In the supranational sphere, it is possible to mention Articles 8 and 9 of the European Convention on Human Rights, Articles 1, 3, 7, 8, and 11 of the Nice Treaty, Article 16 of the Treaty on the Functioning of the European Union, the GDPR, the EU Regulation on the single market for digital services (2022/2065) [13].

In April 2021, the EU Commission also published the well-known proposal for a regulation on AI (COM/2021/206), amended until December 2023.

Before that, in the White Paper on AI (2020), the Commission pointed out the inadequacy of algorithmic decision-making processes, as the machine learning or

deep learning mechanisms, on which these systems are based, operate too fast and on the basis of huge amounts of data. Thus, they produce 'opaque' decisions (black box), as they change over time through the incremental collection of information that cannot be known in advance.

This is the case, for example, of the Chat-GPT platform: its processing of personal data on Italian territory has been temporarily restricted by a precautionary and urgent decision of the competent Independent Authority on March 30th, 2023, for violation of Articles 25, 13, 8, 5, and 6 GDPR. The contested violations concerned: (a) the absence of information on data processing; (b) the absence of a legal basis for the processing of data with the aim of training the algorithms used by the system; (c) the inaccuracy of the processing of the data collected and, therefore, of the information provided; (d) the absence of filters for verifying the age of users (since the developing company had limited the use of the platform to users at least 13 years old, and since minors of that age have not yet acquired—in the opinion of the Authority—an adequate level of self-awareness).

As a consequence of the precautionary measure of the Authority, the platform took actions aimed at supplementing the information and at complying with the further requirements, for instance by ensuring the users the right of opposition even after consent had been given, with the aim of repairing the flaws that the natural opacity of the system may cause.

The critical point, however, is that the behaviour of such systems may be unpredictable by their programmers themselves.

Control activities must therefore be directed at the stages of designing, 'testing', and making operational the software, inhibiting systems that are uncontrollable or capable of becoming uncontrollable.

On the other hand, however, human decision-making processes too are not infallible and, indeed, compared to them, artificial decision-making processes seem immune to certain 'pollution factors' typical of the formers: think to psychological conditioning, corruption, and so on. This immunity is, moreover, sometimes preferred, to the point of opting out of the human decision-making process in favour of the automatic one, even in very delicate contexts: think to the assessment of the patrimonial guarantee for access to credit, or to the adoptive suitability of families.

However, the immunity of automated decisions from the limitations identified for human decisions does not adequately consider that these decisions are the product of the human ones, and may therefore reflect their limitations.

Moreover, big data cannot increase knowledge, limiting themselves at exponentially facilitating its dissemination.

However, the EU's regulatory activism may be interpreted as a proof that the EU's institutions are of the opinion that additional and specific safeguards are necessary.

This activism has been sometimes interpreted as a sign of the transition from the Europe of the market, to that of rights, and finally to that of values, also by promoting the so-called European digital constitutionalism [14–16].

Among the EU member States, Spain adopted a Charter of Digital Rights in July 2021, including the neuro-rights: but it establishes that it does not intend to introduce new rights, being only oriented at making more visible the already existing ones (such

as the digital rights protected by Law 3/2018). In August 2023, it also established the Agency for the Supervision of AI.

As regards non-EU countries, in 2020 Chile proposed a constitutional reform aimed at including the mental integrity among the fundamental human rights enshrined in the Constitution. The present Constitution has been adopted in 1980 and amended several times until 2015. A referendum for the approval of a new Constitution was held in December 2023 with a negative outcome. Thus, the 2020 reform project still refers to the Constitution adopted in 1980. Moreover, a law on the protection of neuro-rights through medical ethics applied to neurotechnologies has been proposed. Also as a consequence of these developments, in August 2023 the Chilean Constitutional Court stated that—according to Article 19 of the Constitution, which protects the psycho-physical integrity of the person and the privacy of his/her data—devices capable of tracking human brain activity for private use must be authorised by the competent health authorities, and the use of the data they collect is subject to the dynamic informed consent of the person concerned: an assent that has to be not only conscious, but also subject to renewal if the purposes for which the data are to be used change from those for which consent had already been given (decision r.n. 105.065–2023) [17].

Table 16.1 reports an analysis of strengths and weaknesses of the affirmative and negative thesis of the need of neuro-rights in order to protect people from the possible threats of AI.

16.4 Conclusions

The choice of possible solutions reflects value options.

Beyond these options, the implementation of the technical knowledge of public decision-makers is essential to ensure a proper relationship between the evolution of technologies and the need to provide an adequate level of protection for people's rights.

The implementation of the technical knowledge of public decision-makers could be realised by increasing the quantity and/or the role of technical bodies and/or data scientists, in supporting the preparatory phase of each regulatory procedure. This activity seems to be essential not only for reducing the 'cognitive deficit' of decision-makers, but also for ensuring a proper relationship between the very rapid evolution of technologies and the need to maintain, or increase, the level of protection of peoples' rights.

The fact that the proposal for an AI regulation established that the Office for AI (recently set up within the Commission by a decision of the same Commission on January 24th, 2024) has to be supported by a scientific group of independent experts, seems to move in this direction.

However, this would remain an inadequate action without a rethinking of the role of public decision-makers. Law should be used more since the design of technologies,

Table 16.1 Analysis of strengths and weaknesses of the affirmative and negative thesis of the need for neuro-rights

Discussion: is there a need for neuro-rights in order to protect people from the possible threats of AI?			
Affermative thesis		Negative thesis	
Strengths	Weaknesses	Strengths	Weaknesses
Specific and more adequate protection of mental integrity, freedom of information and evaluation, mental privacy, personal identity	(1) Excessive protection of new rights as a possible source of weakening of the already existing rights, due to the increase in the level of interpretative contrast. (2) Insufficiency of the establishment of new rights in the absence of the following two conditions: (a) the increase of technical knowledge of public decision-makers; (b) the use of law even in designing new technologies	Enhancing and increasing the adaptability of the existing protections through their extensive interpretation and application	Inadequate response to the following new protection needs: (a) avoid unauthorised mental state disclosure and its use in a predictive way; (b) avoid the so-called 'black boxes', 'cluster effects', tracking activities, social engineering, cognitive bias and/or information asymmetries, loss of control over AI systems

in order to minimise or avoid the threats and enhance the benefits related to their application.

According to the words of Claudio Palomba, law can «discipline the use of new technologies, not their development» [18].

We do not agree with this conclusion, since, on the contrary, we believe that the above mentioned limitations should be considered as the borders to be overcome, not as an absolute hindrance [19–22].

The proposal for a regulation on AI also seems to consider this need, referring to a number of measures in support of innovation, including spaces for regulatory experimentation for AI, functional to carrying out ex ante tests on the development of AI systems in real-life conditions, evaluating their technical-scientific and legal possible effects.

However, some obstacles seem to remain: one material, the other ideological.

The material one concerns the tendency of technologies to develop faster than law. May be this is due to the fact that, although both are the product of human intelligence, the human communities invest more intensively in the new technologies, while the speed of adaptation of law is slowed down by disciplining its production with meticulous rules, aimed at ensuring an adequate level of protection of democracy. This limitation could be governed, though not necessarily avoided, by the above mentioned use of law since the design of new technologies. The different speeds of development, however, should not affect the comparison between human and technological evolution: as long as the human mind remains partially unknown to science, it is unlikely that we will be able to clone people's brains and identities.

However, an additional limit follows this state of the art, i.e. that big data cannot implement autonomously the knowledge of the human community, but only facilitate and exponentially increase its degree of dissemination.

The ideological obstacle concerns the relationship between law and its possible restraining effects on scientific/technological development: can this possibility further slow down the speed of law and limit its role?

The freedom of science must be ensured by accompanying its evolution also through a project-based use of law, to prevent science from slipping into such drifts as to make people its instruments rather than its users.

Obvious, but at high risk of being forgotten.

References

1. Ienca, M.: Neurodiritti, quali nuove tutele per la sfera mentale: tutti i nodi etico-giuridici, in network agenda Digitale, 18 marzo (2021)
2. Ienca, M.: Neurodiritti: storia di un concetto e scenari futuri, in Aa.Vv, Privacy e neurodiritti. La persona al tempo delle neuroscienze. http://www.garanteprivacy.it. Last accessed Jan 2024
3. Gatt, L., Montanari, R., Caggiano, I.A.: Consenso al trattamento dei dati personali e analisi giuridico-comportamentale. Spunti di riflessione sull'effettività della tutela dei dati per-sonali. Pol. dir. **2** (2017)
4. Amato Mangiameli, A.C.: Tecno-diritto e tecno-regolazione. Spunti di riflessione. Riv. fil. dir. (2017)
5. Bellomo, G.: Biometria e digitalizzazione della pubblica amministrazione. In: Ferrara, L., Sorace, D., A 150 anni dall'unificazione Amministrativa Italiana, IV, Civitarese Matteucci, S., Torchia, L. (eds.) La tecnificazione, 1st edn., Firenze University Press, Firenze (2016)
6. Benanti, P.: La dignità della persona al centro dello sviluppo gentile, in Aa.Vv, Privacy e neurodiritti. La persona al tempo delle neuroscienze. http://www.garanteprivacy.it. Last accessed Jan 2024
7. Stanzione, P.: Neurodiritti, I confini della scienza. Corriere della Sera (2021)
8. Trozzi, S.: Il principio della finalità del trattamento dei dati personali alla prova dei recenti sviluppi in tema di intelligenza artificiale: il caso ChatGPT e la neuroprivacy. Federalismi **1** (2024)
9. Bolognini, L.: Mordi la mela? Privacy come diritto del Sè nell'era dell'Intelligenza Artificiale e dell'Augmented Reality. Psiche **2** (2019)
10. Mollo, A.A.: La vulnerabilità tecnologica. Neurorights ed esigenze di tutela: profili etici e giuridici. EJPLT **1** (2021)

11. Gulotta, G., Caponi Beltramo, M.: Neurodiritti: tra tutela e responsabilità. Sistema penale (2021)
12. Cirillo, F.: Ambiguità della "libertà cognitiva" e prospettive di tutela. Consulta on line, 2 (2023). Last accessed Jan 2024
13. Ferri, F.: Transizione digitale e valori fondanti dell'Unione: riflessioni sulla costituzionalizzazione dello spazio digitale europeo. Dir. Ue **2** (2022)
14. Pollicino, O.: Di cosa parliamo quando parliamo di costituzionalismo digitale? Quad. cost. **3** (2023)
15. Pollicino, O.: Costituzionalismo, privacy e neurodiritti. Media Laws (2021)
16. Simoncini, A.: L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà. BioLaw J. **1** (2019)
17. Morelli, C.: Dati neurali, prima sentenza al mondo sulla loro tutela. https://www.altalex.com/documents/news/2023/09/18/dati-neurali-prima-sentenza-al-mondo-sulla-loro-tutela. Last accessed Jan 2024
18. Palomba, C.: Le prospettive della digital transformation tra Politica, Amministrazione e Algoritmi. In: Auby, J.B., De Minico, G., Orsoni, G., L'Amministrazione digitale—Quotidiana efficienza e intelligenza delle scelte, 1st edn., E.S.I., Napoli (2023)
19. Crawford, K.: Né intelligente né artificiale. Il lato oscuro dell'AI, Il Mulino, Bologna (2021)
20. De Heredia Ruiz, I.B.: Inteligencia artificial y neuroderechos: la protección del yo inconsciente de la persona. Aranzadi, Madrid (2023)
21. Hertz, N.: Neurorights—do we need new human rights? A reconsideration of the right to freedom of thought. Neuroethics **16**(5) (2023). https://doi.org/10.1007/s12152-022-09511-0
22. Kellmeyer. P.: 'Neurorights': a human rights–based approach for governing neurotechnologies. In: Voeneky, S., Kellmeyer, P., Mueller, O., Burgard, W. (eds.) Responsible Artificial Intelligence, pp. 412–426. Cambridge University Press (2022)

Chapter 17
Interpretability of Machine Learning Models for Breast Cancer Identification: A Review

Ijaz Ahmad, Alessia Amelio, D. H. Gernsback, Arcangelo Merla, and Francesca Scozzari

Abstract Over recent years, machine learning models have enhanced breast cancer detection, especially in its early stages. Nevertheless, their integration into clinical practices remains limited despite their proven efficacy in early-stage detection among women. However, the results obtained by these approaches are poorly interpretable. This study seeks to demystify early-stage breast cancer detection and boost clinicians' trust in these methods by leveraging eXplainable Artificial Intelligence (XAI). This research underscores the potential of XAI as a foundational step to initiate conversations about adopting supportive AI tools in the clinical sphere. By incorporating these XAI methods, clinicians can better understand why a specific prediction has been made, promoting trust and facilitating more informed decision-making in breast cancer detection and treatment. This study uniquely investigates the potential of advanced XAI techniques to enhance the trustworthiness and reliability of machine learning models, specifically in the early detection and diagnosis of breast cancer. The different XAI approaches are critically reviewed, underlying the current limitations and proposing future work directions.

I. Ahmad
Department of Human, Legal and Economic Sciences, Telematic University "Leonardo da Vinci", 66010 Torrevecchia Teatina, CH, Italy
e-mail: ijaz.ahmad@unidav.it

A. Amelio (✉) · A. Merla
Department of Engineering and Geology, University "G. d'Annunzio" Chieti-Pescara, 65127 Pescara, Italy
e-mail: alessia.amelio@unich.it

D. H. Gernsback
Research Laboratory "Hugo Gernsback", Telematic University "Leonardo da Vinci", 66010 Torrevecchia Teatina, CH, Italy

F. Scozzari
Laboratory of Computational Logic and Artificial Intelligence, Department of Economic Studies, University "G. d'Annunzio" Chieti-Pescara, Pescara 65127, Italy

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_17

17.1 Introduction

In recent years, an explosion of works has been witnessed, proposing machine learning techniques in the field of breast cancer detection, which is still the most frequently diagnosed cancer in women (see, for instance, Chugh et al. [1] and Mridha et al. [2] for recent results). Despite the many results and proposed methods, integrating artificial intelligence techniques in everyday clinical practice is still in an early adoption stage, primarily due to the opaque nature of predictions and the difficulties in explaining the obtained results. Explainability thus plays a vital role in enhancing the understanding and trust in machine learning models.

This paper reviews the most relevant literature on XAI techniques applied explicitly to breast cancer detection. From a total of 30 works, a selection of 18 papers was performed according to (i) relevance in the topic and (ii) completeness and significance of the results. Then, the papers were classified on the XAI method used according to well-known categories found in the literature [3]: (i) perturbation-based (including Shapley Additive exPlanations—SHAP, Local interpretable model-agnostic explanations—LIME, Randomized Input Sampling for Explanation of Black-box Models—RISE), (ii) CAM-based (Class Activation Mapping), (iii) gradient-based, (iv) hybrid, and (v) a last group of other methods which do not fall in the previous categories. For each category, the machine learning method used in each paper is reviewed, together with the obtained results and the proposed XAI techniques to interpret the model result. To the best of the author's knowledge, this is the first work to introduce a similar review in the context of breast cancer identification and interpretation.

The main contribution of the paper is twofold: on the one hand, it describes the current landscape of XAI techniques in breast cancer detection, highlighting the variety and importance of transparent and interpretable models and their limitations; on the other hand, it presents a bunch of different XAI proposals to improve clinicians trust in machine learning models, facilitating integration of supportive AI tools in clinical practice to improve the accuracy of early breast cancer identification and, more in general, the practical implementation and adoption of XAI techniques in healthcare.

The paper is organized as follows. Section 17.2 reviews the papers, describing the machine learning method used, the results' significance, and the specific XAI techniques adopted. In Sect. 17.3, a comparison and discussion of the different XAI methodologies and their limitations is performed. Finally, in Sect. 17.4 conclusions and future work directions are drawn.

17.2 Literature Review

17.2.1 Perturbation-Based Methods

Massafra et al. [4] developed a SHAP-based framework specifically to study breast cancer invasive disease events (IDEs). The research involved 486 breast cancer patients from the IRCCS Istituto Tumori "Giovanni Paolo II" in Bari, Italy. Five-fold cross-validation was executed in twenty iterations. Four machine learning classifiers, Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB), were trained to solve the binary classification problem (IDE vs. no-IDE). Using Shapley values to interpret the contributions of individual features in the prediction model, the authors pinpointed the features driving IDEs. The best classifier was XGB, reaching median Area Under the ROC Curve (AUC) values equal to 93.7 and 91.7% for the 5-year and 10-year IDE predictions, respectively. The primary objective of this framework was to bridge the gap between AI models and their application in real-world clinical scenarios.

SHAP was also used in combination with XGB models by Kumar and Das [5] to pinpoint possible diagnostic markers for breast cancer and by Silva-Aravena et al. [6] as a decision support strategy for breast cancer detection and prevention for healthcare professionals. In the former work, models were trained on a dataset with expression data from peripheral blood mononuclear cells of 252 breast cancer patients and 194 healthy individuals. The dataset was randomly partitioned between 80% training and 20% test sets. The study identified ten critical genes associated with breast cancer progression by integrating SHAP values into the model. The XGB model achieved an accuracy of 94.44%. In the latter work, XGB was used to predict the illness, while SHAP elucidated how different variables influence patient health. The dataset was split into 75% training and 25% testing data. Using ten-fold cross-validation, the XGB algorithm emerged as the most accurate predictor among the tested algorithms, with a train data accuracy of 81.3% and a test data accuracy of 81%. Furthermore, utilizing the SHAP algorithm made it possible to determine the key variables and measure their significance in the prediction.

SHAP was finally adopted in combination with XGB models by Chakraborty et al. [7] to construct data-driven XAI approaches. The study analyzed the association between immune cell composition in the tumor microenvironment and the 5-year survival rates of breast cancer patients based on estimated immune cell composition from bulk RNA-seq data using various cell type quantification methods. A data preprocessing step consisted of encoding categorical information as integer arrays and oversampling to achieve symmetry. Also, hyperparameter tuning using three-fold cross-validation was employed to identify optimal subsets. The proposed approach achieved an accuracy ranging from 96.4% to 100%. Moreover, this explainability method revealed essential details regarding the conditions within the tumor surroundings linked to better patient outcomes and longevity.

Contrary to previous works, Larasati [8], Kaplun et al. [9] and Rafferty, Nenutil and Rajan [10] mainly adopted perturbation-based XAI in combination with shal-

low or deep Artificial Neural Networks (NN) for classification of breast cancer. In particular, Larasati [8] employed Logistic Regression (LR), Gradient Boosted Trees (GBT), and NNs for binary classification of breast cancer and utilized three XAI methods, i.e., SHAP, LIME, and Anchors, for result explanation. The Wisconsin Breast Cancer dataset was employed in this study. The maximum accuracy gained by the GBT and Anchors approaches was 98.57%. An online survey regarding the result interpretation revealed that LIME and Anchors' explanations were challenging for neophytes to grasp. Notably, human explanations aligned better with SHAP. The second approach by Kaplun et al. [9] used Zernike image moments to extract complex features from cancer cell images. The adopted dataset was BreakHis public dataset consisting of histopathology images of breast cancer for automatic prognosis. The dataset was divided into training, validation, and testing, with ratios of 70:15:15. NN models were used for binary classification of malignant and benign cancer cell images. LIME was adopted to explain the obtained results. The proposed approach's recognition rate was 100%. The proposed method can classify cancer cell images at 40× resolution. Finally, the research conducted by Rafferty, Nenutil, and Rajan [10] performed binary classification of breast mammograms as benign and malignant and applied explanation techniques, such as LIME, RISE, and SHAP, to the result. The authors selected mammographic images with the same orientation from the INbreast and DDSM datasets, resulting in 2236 images comprising 1193 benign and 1043 malignant cases. The dataset was divided into training, validation, and testing subsets in a ratio of 95%: 2.5%: 2.5% (2124:56:56). The study employed a publicly accessible Convolutional Neural Network (CNN) architecture, previously utilized for similar tasks but with moderate modifications, achieving an accuracy of 96.43% and an F1 score of 96.42%. The XAI findings revealed a consistent discrepancy between the three methods and the actual medical findings.

17.2.2 CAM-Based and Gradient-Based Methods

A Yolo-based AI model in combination with a gradient-free Eigen-CAM technique was developed by Prinzi et al. [11] for identifying breast cancer in mammograms. The transfer learning technique was used on two publicly accessible full-field digital mammography datasets, i.e., CBIS-DDSM and INbreast. The CBIS-DDSM dataset was randomly divided between 70% training, 15% validation, and 15% test sets. In contrast, the INbreast and proprietary datasets were split into training (80%) and test (20%) sets. The gradient-free Eigen-CAM technique effectively identified all the potentially suspicious regions of interest, even when the predictions were erroneous. The results were promising, especially given the varied nature of the dataset. An Average Precision of 88.5 and 92.2% was obtained for benign and malignant detection, respectively. Saliency maps clarify the model's decision-making and support its integration into clinical tools.

For a similar task, Prodan et al. [12] applied deep learning approaches and vision transformer architectures in combination with Grad-CAM to a publicly available

dataset of mammograms from the Radiological Society of North America. The dataset was split into train/validation sets using a five-fold stratified cross-validation strategy. By adopting synthetic data generation techniques, the classification performance of the models was significantly improved. A ResNet-18 model achieved a maximum accuracy of 94% with an AUC value of 85%. Grad-CAM was used to understand the models' decision-making process.

Finally, Matsuyama et al. [13] constructed an interpretable wavelet-based CNN model for breast density classification using spectral information from mammograms. The authors used a modified ResNet-50 model and proposed algorithms for retrieving image spectra, visualizing network behavior, and quantifying prediction ambiguity. A reliability diagram was used to assess the reliability of the classification model's prediction score, and Grad-CAM was adopted to visualize the basis for the final prediction. The study compared the proposed model with conventional CNN models using image pixel values. Adopting a ten-fold cross-validation, the proposed wavelet model achieved an average accuracy of 92.2% and an AUC of 97.7%.

In the context of gradient-based, Imouokhome et al. [14] employed Integrated Gradient (IG), GradientShap (GS), and Occlusion in combination with a ResNet-50 deep learning model for classifying breast tumors into benign or malignant categories and to elucidate the outcomes. To accomplish this task, the BreakHis dataset was used. The study found that Occlusion visualization surpasses the other two XAI techniques in effectiveness. The proposed approach yielded a validation accuracy of 96.84%.

17.2.3 Hybrid Methods

Hussain et al. [15] developed a visually and mathematically explainable deep learning framework for classifying breast lesions in tomosynthesis lesion images using eight pre-trained CNN models. The study employed two XAI methods for better transparency in a critical clinical context: (i) perceptive interpretability (using Grad-CAM and LIME for visual explanations) and (ii) mathematical interpretability (using t-distributed Stochastic Neighbor Embedding and Uniform Manifold Approximation and Projection for feature clustering). Five-fold cross-validation with stratification was performed, with the validation set comprising 20% images and the train partition containing 80% images. The optimal model recorded AUC values of 98.2% with data augmentation and 96.3% without data augmentation. This research underscored the importance of XAI in understanding AI models, mainly when failures occur.

Also, Khater et al. [16] created a machine learning model specifically designed to classify and interpret breast cancer. To clarify the model's results, the authors utilized three model-agnostic XAI techniques: (i) Permutation Importance (PI), (ii) Partial Dependence Plots (PDP), and (iii) SHAP. Both the Wisconsin Breast Cancer (WBC) dataset and the Wisconsin Diagnostic Breast Cancer dataset (WDBC) were used in this study. Adopting the k-nearest neighbors classifier (KNN), the best-achieved accuracy and precision were respectively 97.7 and 98.2% on the WBC

dataset. Similarly, an NN applied to the WDBC dataset reached 98.6% accuracy and 94.4% precision. Also, it was proved that employing XAI algorithms to analyze breast cancer tumor characteristics significantly improves diagnosis and treatment.

17.2.4 Other Methods

La Ferla et al. [17] endeavored to enhance the early identification of breast cancer and boost clinicians' confidence using XAI that unravels the model's complexity at the neuron level. In particular, the authors reported that "the combination of a ResNet-50 architecture with the Deep Taylor Decomposition methodology has theoretically proved to be a good combination to build a deep learning model and apply backpropagation to it". After evaluating five distinct backpropagation techniques, it became evident that the Deep Taylor Decomposition and the LRP-Epsilon methods yielded superior outcomes.

Also, Amoroso et al. [18] introduced an XAI framework grounded in adaptive dimensional reduction, which pinpoints critical clinical variables for oncological patient profiles and categorizes patients into distinct clusters. The study encompasses data from 267 breast cancer patients. The proposed method ascertains a subspace where hierarchical clustering capitalizes on patient distances to deduce the most fitting categories. Pertinent data underscores that molecular subtype (vs. therapy combination) is paramount for effective clustering. The Adjusted Rand Index (ARI) between clustering results and molecular subtypes (vs. therapy combinations) was 0.6 and 0.4, respectively, for a number k=4 of clusters. The resilience of prevailing therapeutic guidelines was examined, revealing a marked association between unsupervised patient profiles and molecular subtypes.

In the context of deep learning, Dong et al. [19] explored the decision-making process of a deep model in lesion classification by introducing a new "region of evidence" (ROE) and incorporating XAI techniques. The study used coarse and fine regions of Interest (ROIs) to examine the robustness of the deep learning model and the impact of lesions and peripheral tissues on classification outcomes. DenseNet-121 classified breast lesions in 2D grayscale ultrasound images using a dataset of 785 2D breast ultrasound scans from 367 women. The coarse and fine ROI models produced AUCs of 89.9% and 86.9%, respectively. The model achieved 88.4% accuracy, 87.9% sensitivity, and 89.2% specificity with coarse ROIs and 86.1, 87.9%, and 83.8% with fine ROIs. This ROE-based metric offered a more transparent insight into how AI interprets images, enhancing comprehension for both physicians and patients.

Similarly, Zhang, Vakanski, and Xian [20] introduced an advanced, interpretable deep network diagnostic system (BI-RADS-Net-V2) for binary classification of tumors from breast ultrasound images. Incorporating medical insights from the Breast Imaging Reporting and Data System (BI-RADS) ensured dependable and streamlined interpretations for medical practitioners. Notably, BI-RADS-Net-V2 adeptly differentiated between malignant and benign tumors, delivering both semantic and quantitative rationale for its determinations. Evaluations were conducted on a dataset

of 1,192 breast ultrasound images split into 80 and 20% for training and testing using a five-fold cross-validation approach. The proposed BI-RADS-Net-V2 gained an accuracy of 88.9%.

Finally, to detect cancer through mammograms from the CBIS-DDSM dataset, Bouzar-Benlabiod et al. [21] integrated the U-Net deep learning model with a Case-Based Reasoning (CBR) system to enhance classification accuracy. The proposed approach leverages deep learning models for precise mammogram segmentation and CBR for explainable, accurate classification. The tested method demonstrated high effectiveness, achieving an accuracy of 86.71% and a recall of 91.34%. All the previously described articles are summarized in Table 17.1.

17.3 Discussion and Current Limitations

From Table 17.1, it is worth noting that perturbation-based methods, such as SHAP, LIME, and RISE, are the most used ones for machine learning classifiers (with the only exception of Rafferty, Nenutil, and Rajan [10]). By contrast, gradient-based methods, such as Integrated Gradients, its extension GradientShap and SmoothGrad, and CAM-based methods, such as Grad-CAM and its variants, have been introduced explicitly in the literature for explaining deep networks, e.g., CNNs.

In particular, perturbation-based methods, i.e., SHAP, LIME, and RISE generate a set of perturbations on the original data to produce a new set of predictions, which are then compared with the initial results to assess the relevance of different portions in the original data. Accordingly, they can also be adopted for machine learning models that do not provide information on gradient or feature maps, such as NB, SVM, LR, and RF. In fact, NB uses the Bayes theorem for computing, for each data instance, the probability of belonging to a given class. RF combines the classification of different decision trees, considered weak classifiers, for detecting the final class. SVM aims to identify the hyperplane's parameter values, optimally separating the instances belonging to different classes. Finally, LR seeks to estimate the parameters of a logistic model. By contrast, gradient-based methods back-propagate the gradient and its variants to the original input image for detecting the pixels or regions that have the most significant impact on the change of prediction. Also, CAM-based methods combine the forward feature maps using backpropagation gradients adopted as weights to separate objects from the background.

However the choice of a specific XAI method depends also on finding a balance between the model accuracy and the easiness of interpretability. As pointed out by Larasati [8], LIME and Anchors' explanations can be more challenging to interpret since they require a better understanding of how a locally approximated model behaves. On the contrary, SHAP values indicate how each feature contributes to the model's output, which is easier for clinicians since it carries out a global interpretation of the importance of each feature, which is valid across all the instances. This could explain the high number of works employing SHAP-based XAI. The hybrid methods try to fill this gap by exploiting a combination of techniques, such as Hussain

Table 17.1 XAI methods for breast cancer identification

Work	AI technique	XAI category	XAI method	Results (%)
Massafra et al. [4]	NB, SVM, RF, XGB	Perturb.	SHAP	AUC of XGB 93.7/91.7 (5y/10y)
Kumar and Das [5]	XGB	Perturb.	SHAP	Acc. 94.44
Silva-Aravena et al. [6]	XGB	Perturb.	SHAP	Acc. 81
Chakraborty et al. [7]	XGB	Perturb.	SHAP	Acc. 96.4-100
Larasati [8]	LR, GBT, NN	Perturb.	SHAP, LIME, Anchors	Acc. of GBT and Anchors 98.57
Kaplun et al. [9]	NN	Perturb.	LIME	Acc. 100 (40× img.)
Rafferty et al. [10]	CNN	Perturb.	LIME, RISE, SHAP	Acc. 96.43; F1 96.42
Prinzi et al. [11]	Yolo-based models	CAM	Gradient-free Eigen-CAM	AP 88.5/92.2 (ben./mal.)
Prodan et al. [12]	ResNet-18	CAM	Grad-CAM	Acc. 94; AUC 85
Matsuyama et al. [13]	Modified ResNet-50	CAM	Grad-CAM	Acc. 92.2 AUC 97.7
Imouokhome et al. [14]	ResNet-50	Grad.	IG, GS, Occlusion	Acc. 96.84
Hussain et al. [15]	Pre-trained CNNs	Hybrid	Grad-CAM, LIME	AUC 98.2/96.3 (aug./no aug.)
Khater et al. [16]	KNN, NN	Hybrid	PI, PDP, SHAP	Acc. 97.7, 98.6 Prec. 98.2, 94.4
La Ferla et al. [17]	ResNet-50	Other LRP-Epsilon	Deep Taylor Decomp.	Detecting areas related to class. of benign or malignant cancer
Amoroso et al. [18]	hier. clustering	Other	Adaptive dim. reduction	ARI (k=4) 60/40 (mol. subt./ther.)
Dong et al. [19]	DenseNet-121	Other	ROE-based technique	AUC 89.9/86.9 (coarse/fine)
Zhang et al. [20]	BI-RADS-Net-V2	Other	BI-RADS-Net-V2	Acc. 88.9
Bouzar-Benlabiod et al. [21]	U-Net	Other	CBR system	Acc. 86.71; Rec. 91.34

et al. [15], which emphasize the importance of combining visual and mathematical explainability, integrating both Grad-CAM and LIME.

From all the aforementioned, the main limitations of the proposed XAI methods are still in the scarce attention to provide a user-friendly interface which should enable clinicians to directly interact with the XAI models to improve understanding and trust. Thus the right balance between accuracy and interpretability of the XAI method is still the main challenge, which may be addressed by fostering a more strict collaboration between computer scientists and clinicians. Another limitation is the variety of XAI techniques, which is simultaneously a strength since it forces clinicians to deal with different XAI methods, which is more challenging and does not promote confidence in the model. Different XAI techniques offer distinctive methodologies for interpreting the inner workings of models, allowing healthcare professionals to select the approach that best suits their specific needs and the characteristics of the available data. However, this implies the need for a thorough understanding of the different techniques and their applications in the diagnostic context and the demand for specialized training and skills by healthcare professionals and XAI experts. Also, the simultaneous use of different XAI techniques can lead to overlapping and redundant information provided by other models. This can complicate the interpretation of results and make it challenging to identify diagnostically relevant information. Ultimately, the clinical decisions that result from XAI techniques may be influenced by the subjective and professional judgments of healthcare practitioners, thereby introducing the possibility of divergent interpretations of model outcomes.

17.4 Conclusions and Future Work

This paper provided a critical review of different XAI methods adopted in the literature for the interpretability of breast cancer identification. The most used ones are perturbation-based, with particular reference to SHAP, since its interpretability output is more straightforward for clinicians to understand. This is followed by the hybrid methods combining both visual and mathematical explainability.

From the studies conducted, multiple research topics could be developed in the future.

In particular, given the variety of XAI methods, it would be helpful to develop predictive models that automatically identify the most effective XAI techniques, as well as to explore new approaches and tools that effectively integrate and visualize the explanations provided by different XAI techniques in a single user-friendly user interface. Such standardization could bring many benefits, not only for easiness of interpretability but also to further the development of standardized evaluation metrics for XAI methods, which in turn can facilitate the benchmarking and comparison across different studies. Also, a well-designed and intuitive user interface can enable clinicians to quickly access relevant information and efficiently interpret the results of XAI models.

In general, exploring strategies for developing XAI-based intuitive and easy-to-use interactive tools for health care providers, incorporating dynamic visualization and contextualized interaction features, would be advantageous. One direction could be to adopt Advanced Visualization Techniques that can be used to highlight the "most relevant" or crucial attributes for model predictions. For example, SHAP or LIME attribute maps can show which data features contribute most to a specific prediction, making it easier even for those who are not machine learning experts to understand the model's reasoning and increase confidence in such tools. In addition, the relative importance of each input variable for the model predictions can be highlighted, and their impact on the model predictions can be visualized through simple graphs or other graphical representations to provide health professionals with a clear overview of the most influential variables in model decision-making. In this approach, visualizations should be supplemented with contextualized explanations to facilitate understanding. Equally helpful, in this regard, would be the ability to make health care providers easily customize both user interface settings and preferences to suit their needs and visualization preferences by enabling clinicians and patients to visually examine and understand the model's predictions and the underlying reasoning behind them (e.g., by allowing users to select specific variables to be displayed, change model parameters, and customize the appearance of the user interface). The user interface could incorporate decision support tools, which assist healthcare professionals in analyzing model predictions and making informed decisions regarding diagnosis and treatment. They can facilitate more effective detection of potential errors or biases in models or risks of overfitting by enabling medical staff, also through their active involvement in interpreting predictions, to make corrections and improvements to models to reduce the risk of misdiagnosis or inappropriate treatment.

Another aspect to be explored is the establishment of metrics and guidelines for implementing and evaluating XAI techniques to improve the consistency and reliability of their use in different studies and clinical applications. These metrics would simplify the validation of studies, as they would reduce the time required for this process and the variability in the results obtained from different implementations of XAI techniques, resulting in a reduction in the resources needed for validation and an improvement of the overall efficiency of the evaluation. This could involve developing standardized procedures for data preparation, model training, and performance evaluation, perhaps comparing the performance of different XAI techniques. In addition, it may be necessary to define new uniform metrics, less sensitive to variation in data representation, to assess the effectiveness and interpretability of XAI models. These metrics should be carefully balanced against a potential loss of flexibility and adaptability. On the one hand, healthcare providers and researchers may hesitate to explore new methodologies and approaches if they do not conform to existing standardized procedures, thus limiting the scope for innovation and exploration of new ideas and techniques in the field of AI. On the other hand, the scope of metrics should be narrowed to fit the specific clinical contexts of diagnosis and patients' individual needs.

References

1. Chugh, G., Kumar, S., Singh, N.: Survey on machine learning and deep learning applications in Breast Cancer Diagnosis. Cogn. Comput. **13**(6), 1451–1470 (2021). https://doi.org/10.1007/s12559-020-09813-6
2. Mridha, M.F., Hamid, M.A., Monowar, M.M., et al.: A comprehensive survey on deep-learning-based breast cancer diagnosis. Cancers **13**(23), 6116 (2021). https://doi.org/10.3390/cancers13236116
3. Wang, H., Wang, Z., Du, M., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 111–119. IEEE Computer Society, Los Alamitos, CA, USA (2020). https://doi.org/10.1109/CVPRW50498.2020.00020
4. Massafra, R., Fanizzi, A., Amoroso, N., et al.: Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front. Med. **10** (2023). https://doi.org/10.3389/fmed.2023.1116354
5. Kumar, S., Das, A.: Peripheral blood mononuclear cell derived biomarker detection using eXplainable Artificial Intelligence (XAI) provides better diagnosis of breast cancer. Comput. Biol. Chem. **104**, 107867 (2023). https://doi.org/10.1016/j.compbiolchem.2023.107867
6. Silva-Aravena, F., Núñez Delafuente, H., Gutiérrez-Bahamondes, J.H., et al.: A hybrid algorithm of ML and XAI to prevent breast cancer: a strategy to support decision making. Cancers (Basel) **15**, 1–18 (2023). https://doi.org/10.3390/cancers15092443
7. Chakraborty, D., Ivan, C., Amero, P., et al.: Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer. Cancers (Basel) **13**(14), 3450 (2021). https://doi.org/10.3390/cancers13143450
8. Larasati, R.: Explainable AI for breast cancer diagnosis: application and user's understandability perception. In: International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1–6. IEEE. https://doi.org/10.1109/ICECET55527.2022.9872950
9. Kaplun, D., Krasichkov, A., Chetyrbok, P., et al.: Cancer cell profiling using image moments and neural networks with model agnostic explainability: a case study of breast cancer histopathological (breakhis) database. Mathematics **9**(20) (2021). https://doi.org/10.3390/math9202616
10. Rafferty, A., Nenutil, R., Rajan, A.: Explainable artificial intelligence for breast tumor classification: helpful or harmful. In: Reyes, M., Henriques Abreu, P., Cardoso, J. (eds.) iMIMIC-2022. LNCS, vol. 13611, pp. 104–123. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-17976-1_10
11. Prinzi, F., Insalaco, M., Orlando, A., et al.: A yolo-based model for breast cancer detection in mammograms. Cogn. Comput. **16**, 107–120 (2023). https://doi.org/10.1007/s12559-023-10189-6
12. Prodan, M., Paraschiv, E., Stanciu, A.: Applying deep learning methods for mammography analysis and breast cancer detection. Appl Sci **13**(7), 4272 (2023). https://doi.org/10.3390/app13074272
13. Matsuyama, E., Takehara, M., Takahashi, N., et al.: A breast density classification system for mammography considering reliability issues in deep learning. Open J Med Imaging **13**, 63–83 (2023). https://doi.org/10.4236/ojmi.2023.133007
14. Imouokhome, F.A.U., Ehimiyein, O.G., Chete, F.O., et al.: Diagnosis and interpretation of breast cancer using explainable artificial intelligence. NIPES J Sci Technol Res **5**(2), 102–123 (2023). https://doi.org/10.5281/zenodo.8014197
15. Hussain, S.M., Buongiorno, D., Altini, N., et al.: Shape-Based breast lesion classification using digital tomosynthesis images: the role of explainable artificial intelligence. Appl Sci **12**(12), 6230 (2022). https://doi.org/10.3390/app12126230
16. Khater, T., Hussain, A., Bendardaf, R., et al.: An explainable artificial intelligence model for the classification of breast cancer. IEEE Access (99):1–1 (2021). https://doi.org/10.1109/ACCESS.2023.3308446
17. La Ferla, M., Montebello, M., Seychell, D.: An XAI approach to deep learning models in the detection of Ductal Carcinoma in Situ, pp. 1–12 (2021) . arXiv:2106.14186

18. Amoroso, N., Pomarico, D., Fanizzi, A., et al.: A roadmap towards breast cancer therapies supported by explainable artificial intelligence. Appl Sci 11(11), 4881, 1–17 (2021). https://doi.org/10.3390/app11114881
19. Dong, F., She, R., Cui, C., et al.: One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. European Radiol **31**, 4991–5000 (2021). https://doi.org/10.1007/s00330-020-07561-7
20. Zhang, B., Vakanski, A., Xian, M.: BI-RADS-NET-V2: a composite multi-task neural network for computer-aided diagnosis of breast cancer in ultrasound images with semantic and quantitative explanations. IEEE Access **11**, 79480–79494 (2023). https://doi.org/10.1109/access.2023.3298569
21. Bouzar-Benlabiod, L., Harrar, K., Yamoun, L., et al.: A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. Comput Biol Med **163**, 107133 (2023). https://doi.org/10.1016/j.compbiomed.2023.107133

Chapter 18
Constitutional Challenges and Regulatory Framework: Will the EU's Artificial Intelligence Act Ensure Adequate Protection of Fundamental Rights and Democracy?

Pietro Masala

Abstract The development of data-driven technologies and particularly of artificial intelligence entails, together with opportunities, serious risks for the individuals and society, including discrimination, manipulation, adverse impact on fundamental rights and democracy. Liberal-democratic societies are therefore called to address major constitutional challenges, by providing adequate legal responses, which should ensure a responsible use of new technologies, consistent with the overriding need to safeguard constitutional values. The need for a regulatory framework is especially felt in Europe, within both the Council of Europe and the European Union, although approaches and tools are partially different. A more advanced awareness of present challenges is shown by recent CoE non-binding texts, based on a human rights-centered approach, whereas the EU, since the 2016 General Data Protection Regulation, has adopted a strategy based on a twofold approach, aiming to combine the development of digital economy with the safeguard of constitutional values. This paper presents a critical analysis of the AI Act, a fundamental part of such strategy, and regards it from a Constitutional Law perspective, by assessing its capacity to ensure adequate protection of fundamental rights and democracy. This requires to consider the troubled history of the act, starting from the European Commission's proposal, which was followed by long and difficult negotiations. The analysis is intended to highlight the act's weaknesses and necessary changes.

P. Masala (✉)
Department of Legal and Social Sciences, "D'Annunzio" University of Chieti-Pescara, Pescara, Italy
e-mail: pietro.masala@unich.it

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_18

18.1 Introduction: Constitutional Challenges in the Era of Big Data and Artificial Intelligence

The development of data-driven technologies and especially of Artificial Intelligence (AI) entails, together with opportunities, serious risks for both the individuals and society. In particular, AI applications, which are in general based on machine learning systems fueled by an extraordinary amount of information made available by an impetuous process of "datafication" of reality and human lives and by the refinement of data extraction and processing techniques, are increasingly being used for decision-making both in the public and private sectors, across a large variety of areas including, among others, work, education, social welfare, justice, and law enforcement. If their use may imply a gain in terms of time and efficiency—or this is the alleged justification for their diffusion—the quality of the decisions which are thus made is in many respects controversial. By contrast, it is certain that use can have a significant adverse impact on the existences of the persons exposed, on their self-determination and life conditions, fundamental freedoms and rights, even coming to affect, due to its increasing spread and pervasiveness, society and the very functioning of democracy. This raises serious concerns and questions about the compatibility of such a use of AI systems with the safeguard of modern constitutional values [3, 13, 17–19, 23].

Risks are inherently connected to the specific characteristics of big data, machine learning and AI: namely, to the statistical (rather than logical-deductive) approach which is used to establish connections among data available in enormous amounts and contained in an extraordinary variety of datasets and so make inferences; to the opacity of the algorithmic decision-making which is based on the processing and use of such data; to the possible use of biased datasets, which, whether producers and users are aware or not of biased inputs, entails biased results and decisions affecting individuals and groups [6, 14, 23]. Such characteristics in themselves give rise to heavy risks of manipulation and discrimination, which are, in this sense, structural, and whose harmful repercussions on fundamental rights and democracy are evident. Moreover, some AI systems entail specific and higher risks, due to which they must be considered irreconcilable with individual freedoms, democracy and the rule of law [13, 20, 23]. This is the case, in particular, of predictive optimization systems (POS), intended to supposedly predict human behavior and make decisions affecting the individuals and society. Even when they are not based on pseudo-scientific theories (like emotion recognition systems: ERS), it is certain that such artifacts, which de facto claim to treat human persons, groups and society as things, have inherent political qualities, as they implicitly deny and are incompatible with individual self-determination, which is the basis of liberalism and modern constitutionalism [20, 21].

Data processing in the AI era has therefore collective implications which must be duly taken into account by policy makers and legislators, if these want to effectively safeguard the essential values enshrined by national Constitutions and international Charters [9–11]. In other terms, liberal-democratic societies are called to address major constitutional challenges [13, 15], by providing adequate legal responses,

which should ensure a responsible use of big data, algorithmic decision-making and AI, consistently with the overriding need to safeguard those values.

In this perspective, it is argued that the described characteristics of AI applications and the extent of the risks involved should lead policy-makers and legislators to the adoption of a precautionary approach [17, 19], as the first choice (about admitting, regulating, or banning certain systems and practices) is decisive [20, 23]. A regulatory framework should therefore include prohibition of all systems whose effects are not certain or cannot be understood and foreseen satisfactorily, at least in areas which are crucial for the protection of fundamental rights, democracy and the rule of law. Otherwise, it would only legitimize and normalize the use of unsafe systems.

Risks are increased by both the impetuousness of technological advances and the lack of proper regulation: this is a difficult goal to achieve, not only because an effective legal protection system should keep pace with such advances but also because its adoption and implementation must deal with conflicting interests and powers of different kinds. In fact, it is frequently seen as an obstacle both by private operators and especially big technological companies, interested in maximizing their profits without any legal limitations, and by governments, which may be either concerned for the possible loss of competitiveness at the global level which could be caused by a strict regulation or reluctant to renounce to possible uses for law enforcement and national security. The safeguarding of individual rights and democracy must thus cope with two great enemies: firstly, the new private power of Big Tech (grounded on unprecedented concentrations of capital and knowledge, which were made possible precisely by the lack of proper regulation); secondly, public power, which has too often shown to be weak and indulgent in its relationship with the first one and may, for its part, be seduced by illiberal and neo-authoritarian temptations. If that safeguarding has to be achieved as a primary objective, then both powers should be equally limited by law, consistently with the original purpose of constitutionalism [13, 17, 18, 23]. In this context, legal scholars should be aware of present risks, so that, on one side, when regulation is lacking, they can identify and recommend proper responses *de iure condendo*, also contributing to increase the awareness of policy makers and citizens; and so that, on the other side, *de iure condito*, they can critically analyze the responses given by legislators, in order to detect shortcomings and highlight necessary changes.

18.2 Legal Responses in Europe: CoE and EU's Approaches

The need for a regulatory framework is especially felt in Europe [5, 16]. Both the Council of Europe (CoE) and the European Union (EU) have addressed the constitutional challenges raised by the development of data-driven technologies and AI, by recently adopting legal texts of various kinds and forces. This is not surprising. Firstly,

the principles enshrined in the highest legal acts (Treaties and Charters) of those organizations epitomize the constitutional values which have been widely shared by the European States since the end of World War II (even more since the end of the Cold War, although the consensus around the postwar "*ius publicum europaeum*" has recently been questioned by the worrying rise of populism and "illiberalism" in several countries) and are currently exposed to the risks raised by the technological revolution. Secondly, because, in order to hope to effectively tackle those risks, responses must be agreed and given at the international or supranational level.

However, the approaches and tools adopted in face of common challenges within the two distinct legal systems are partially different. A more explicit awareness of risks for constitutional values is shown by some recent legal documents adopted within the CoE, which are based on a human rights-centered approach, such as the modernized Convention for the Protection of Individuals with regard to the Processing of Personal Data (2018) and especially the Guidelines on the Protection of Individuals with regard to the Processing of Big Data (2017) and on AI and Data Protection (2019). These texts require or recommend specific consideration of ethical and constitutional values, also as enshrined by international Charters, as far as it concerns risk assessment, thus proposing for this a values-oriented and participatory model [8–11]. However, the CoE does not have relevant legislative powers, but essentially provides a system of judicial protection of fundamental rights. As its guidelines are not binding, their capacity to govern the impetuous ongoing transformations is limited. On the other hand, they may represent important references and hopefully influence national legislation and EU law, also according to dynamics of cross-fertilization [11].

By contrast, the EU institutions are conferred legislative powers in extended and relevant areas by the European Treaties and can exercise them by adopting binding acts, such as regulations and directives, in compliance with the principles of EU primary law, including the Charter of Fundamental Rights. In fact, by exercising such powers and namely the competence on the internal market (Art. 114, Treaty on the Functioning of the EU), they have, since the adoption of the General Data Protection Regulation (Regulation 2016/79: GDPR), pursued a general strategy concerning new data-driven technologies, claiming to be based on a "balanced approach," i.e., on the twofold concern to combine the development of digital economy with the protection of fundamental rights and European constitutional values. We already analyzed the GDPR's model of risk assessment and pointed out its limits, if compared to the CoE's model, suggesting that the CoE's human rights-centered general approach may explain the differences; but we also noted that, even though the data protection impact assessment as described in Art. 35 GDPR is characterized by quite generic references to fundamental rights and participation, it could be interpreted and implemented consistently to a broader view of the impact of data processing, including assessment of collective risks and effective participation of stakeholders [8, 11]. It must also be recalled that the list of the data subjects' rights enshrined in the GDPR includes the new "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her," though with important exceptions (Art.

22: a general provision which can evidently be referred to AI systems too). Moreover, awareness of collective risks and recommendations to adopt a values-oriented and participatory approach in their assessment have been present also within the EU for a quite long time, as made evident by some resolutions of the European Parliament (EP; e.g., Resolution of 14 March 2017 on *Fundamental rights implications of big data*) and—we will see examples in next section—by several opinions of the European Data Protection Supervisor (EDPS), an independent institution responsible (currently under Art. 32, Regulation 2018/1725) "for advising Union institutions and bodies and data subjects on all matters concerning the protection of personal data" [11].

The European Commission (EC) chaired by Ursula von der Leyen (2019–2024) has implemented the above mentioned strategy, by presenting proposals for further binding acts on specific fields, intended to complement the GDPR, which have subsequently been debated and, in some cases, passed (among them, also the so-called Digital Service Act and Digital Markets Act). What we propose here is to provide a critical analysis of the so-called Artificial Intelligence Act (AI Act), a fundamental part of that strategy, and namely to regard it from a Constitutional Law perspective, by assessing its capacity to ensure adequate safeguard of fundamental rights and democracy. This requires to consider the troubled history of the act, as the draft proposal for a Regulation on AI presented by the EC on 21 April 2021 was followed by long and difficult negotiations among the EU's co-legislative bodies. The analysis is intended to highlight the act's approach and its weaknesses: thus, after an overview of the relevant aspects of the draft proposal, it will consider the negotiations' results, in order to further highlight necessary changes.

18.3 The EU Artificial Intelligence ACT and Constitutional Values

18.3.1 The Draft Proposal: Approach and Shortcomings

Due to the limited extension of this contribution and the complexity of the AI Act, we cannot develop here a detailed analysis. Nonetheless, we can describe the main features of the EC's proposal in order to point out its main shortcomings concerning the protection of constitutional values. This will later allow us to synthetically consider the text resulting from the subsequent negotiations, so as to assess to which extent those shortcomings persist or further reasons for criticism can be found.

The proposal—COM(2021)206—was based on Art. 114 TFEU, i.e., on the general competence on internal market, entailing the power to adopt harmonization measures in this area. The related Explanatory memorandum explicitly connected the draft to the EC's general strategy which claims to be based on a "balanced approach": in fact, it stressed that "the same elements and techniques that power the socio-economics benefits of AI can also bring about new risks or negative consequences

for the individuals or the society"; and that the proposal "s[ought] to ensure a high level of protection" for fundamental rights affected by AI and "aim[ed] to address various sources of risk through a clearly defined risk-based approach," which "does not create unnecessary restrictions to trade, whereby legal intervention is tailored to those concrete situations where there is a justified for concern." In other terms, the risk-based approach was intended to be proportionate, implying different rules (prohibitions or specific obligations for development, placement on the market and use) depending on the level of risk associated with different kinds of AI systems. The question is: is this approach, as reflected in the proposal (and later in the act), actually "balanced" and always proportionate? Or, more precisely: is it adequately balanced? Does it ensure adequate safeguarding of constitutional values? Before examining its provisions and trying to answer, it is worth recalling that the draft was preceded by a White Paper on AI, also released by the EC; and that in its opinion n. 4/2020, on this document, the EDPS had welcomed numerous references to a European approach on AI, grounded in EU values and fundamental rights, but also noted that the notion of risk impact seemed "too narrowly defined." In fact, it considered that, "beside 'the impact on the affected parties', the assessment of the level of risk of a given use of AI should also be based on wider societal considerations, including the impact on the democratic process, due process and the rule of law, the public interest, the potential for increased general surveillance, the environment and (concentrations of) market power."

The subsequent proposal was not equally explicit at the moment of setting out its risk-based approach in detail and specifying the tools to address different risks. To a first approximation, this approach consisted (and that has not changed as a result of negotiations) of a combination of banning, obligations, accountability, self-regulation, and soft law, where banning is the exception. In concrete, Title II of the proposal banned those AI systems whose use was considered unacceptable as contravening Union values, namely by violating fundamental rights (Art. 5). Prohibitions were established without exceptions only for manipulative systems (using subliminal techniques or exploiting vulnerabilities of a specific group of persons; but with a harm requirement reducing protection effectiveness) and social scoring, whereas real-time and remote biometric identification systems were prohibited except for specific law enforcement purposes if accompanied by an independent authorization regime.

With regard to high risk systems, that pose significant risks to the health, safety and fundamental rights of persons, Title III contained classifications rules and a list (Art. 6, as complemented by two annexes); established the EC would be empowered to update this list by adopting delegated acts (i.e., more easily and faster than by the co-legislative process), considering *inter alia* the "risk of adverse impact on fundamental rights (Art. 7); governed "testing procedures" and a "risk management system" (Art. 9), as well as "data governance" (for training, validation and testing of data sets), including "examination in view of possible biases" (Art. 10). More specifically, Title III would apply to both AI systems that are products or components of products already covered by existing EU safety harmonization legislation (listed in annex II) and to standalone AI systems, as specified in annex III for use in eight

areas (biometric categorization, both remote, as in Title II, and applied "post" the event; management and operation of critical infrastructure; educational and vocational training; employment, worker management and access to self-employment; access to and enjoyment of essential services and benefits; law enforcement; migration, asylum and border management; administration of justice and democracy). It was established that, subject to EP or Council veto, the EC could add sub-areas within these areas, if the application poses a similar risk to an existing on-scope application, but could not add new areas entirely (Arts. 7 and 73). For both subcategories, this title contained a list of requirements (Chapter II), connected to obligations of regulated actors, and especially providers, who must undergo conformity assessment (Chapter III). This would be concretely based on harmonized standards adopted by European Standardization organizations [22], and, consistently with a general principle of accountability (which, as is well known, is central in the GDPR), a self-assessment. Title IV established specific transparency obligations for AI users and providers, applying to some systems (systems intended to interact with natural persons: "bots"; emotion recognition and biometric categorization systems: systems generating or manipulating image, audio, video: "deep fake"): essentially disclosure obligations, requiring to inform the persons exposed, but with significant exceptions, namely for legally authorized crime prevention purposes.

A more exhaustive analysis cannot be carried out here, but we can consider the reception of the proposal. Some scholars welcomed its general approach, insofar as they deemed it as balanced in reconciling the aim to provide a framework for the development of digital economy with the aim to safeguard fundamental rights and constitutional values [4]; or at least proportionate (as it took into account the complexity and variety of AI systems in classifying them according to their different risk level and subsequently establishing specific requirements and obligations) and reasonably flexible—e.g., by conferring the EC the power to update the list of high-risk systems by delegated acts and in general for the importance given to conformity standards and soft law. Nevertheless, criticism was also widely expressed by those who, while considering the flexibility and proportionality of the approach positively, "demystified" the proposal, by highlighting its shortcomings, and expressed concern for its capacity to safeguard individual self-determination, fundamental rights and democracy, despite the declared intention to ensure respect of European constitutional values [22]. Critics concerned, in general, the concrete definition of the risk-based approach, as, in accordance with the principle of accountability enshrined by the GDPR and further than that, the EC shaped a protection system mainly built around the operator (provider or user, as seen). It was underlined that, by contrast, the person subject to the effects of an AI application was not provided with specific legal remedies in addition to those contained in Art. 22 GDPR; and that the proposal could be considered as grounded on a consumer-centric rather than human-centric approach, also because protection tools (such as the quality system based on the conformity with European standards and on self-assessment) manifestly replicated those previously introduced in general EU legislation concerning the safety of some commercial products [12]. Concerns were also expressed with regard to the minor role of public independent authorities and to the conformity assessment made by the

operators, based on standards established by private companies, whereas a strong control would be required by the care of primary public interests (constitutional values), which would be seriously affected by the use of not prohibited and though essentially high-risk or unacceptable AI systems [12, 20, 22]. Further shortcomings are the limited territorial scope (the act would only apply to AI systems placed on the Union market); and the risk of preemption of national legislations protecting rights and values more effectively, in the name of harmonization within the internal market [22]. Quite remarkably, criticism was also expressed in the joint opinion of the European Data Protection Board and of the EDPS on the proposal (n. 5/2021), underlining the absence of remedies and rights directly enforceable by the person subjected to the use of AI systems and the lack of a clear coordination with the GDPR protection system; and by the Italian data protection authority, in its memorandum submitted to the Chamber of Deputies on 9 March 2022, because of the limited role of EU and national supervision authorities and the consequent risk of weakening fundamental rights protection.

18.3.2 After Negotiations: Persisting Weaknesses and New Concerns

The EC's draft proposal, according to the co-legislation process applied in the area of internal market, had to be discussed by the EP and by the Council. Negotiations were long and hard, as they resented both the lobbying activity of the industry and the resistance of some national governments (represented within the Council) to stricter regulation. The EP introduced some improvements of the protection system (by extensively banning biometric surveillance and predicting policies; by introducing an obligation to perform a fundamental rights impact assessment for deployers of high-risk systems), but neither questioned the general approach (centered on the operators' accountability rather than on making legal remedies directly available to the person exposed to the effects of an AI system) nor fully banned all controversial practices, maintaining exceptions also for biometric identification systems, namely in the migration context [1]. Difficult trilateral negotiations followed and finally resulted in a deal reached on 8 December 2023, a general political compromise to be specified in significant aspects [7]. Further technical meetings were therefore necessary to produce a consistent version of the AI Act, which was adopted by the Council on 2 February 2024 and is now set for being finally adopted by the EP.

Serious reasons for criticism persist, both because significant shortcomings which were already present in the initial proposal remain and because this has been, in other respects, made worse. A general weakness is the limited use of prohibition, accompanied by exceptions and requirements reducing the scope and effectiveness of the protection against dangerous practices: with the effect to legitimize these rather than protecting the individuals and society, especially vulnerable groups. The

devil is often in the details, which often entail discrimination: a worrying "two-tiered approach to fundamental rights, which deems migrant people and already-marginalized people less worthy of protection" was present in the proposal and persists [7]. Among the main shortcomings, it is worth mentioning the case of ERS, that claim to infer emotion from biometrics: their use is only banned "in the areas of workplace and education institutions" (subject to an unclear "safety" exception), not in contexts such as law enforcement and migration, and they are classified as high-risk systems. And this, despite serious concerns about their scientific validity; despite the fact their intrusive nature would increase the imbalance of power between the person concerned and the public authority, in addition to definitely implying objectification of the person; despite the fact that the harm to the persons concerned (vulnerable persons, like asylum seekers), and namely the adverse impact on their rights to privacy, autonomy and dignity, is not a risk but a certainty [2]. Moreover, a new "structural loophole" was introduced by the deal reached in December 2023: the addition of a 'filter' into the classification system of high-risks applications in Art. 6, i.e., the inclusion of broader criteria offering developers avenues to exempt themselves from obligations to protect people's rights, by unilaterally deciding if they believe the system is high-risk [7].

18.4 Final Remarks

Over the last decade, the EU has represented itself as being pursuing the goal of building a Digital Single Market within which setting high standards in data protection should favor the trust of data subjects and thus economic growth based on the fair implementation of data-intensive technologies, including AI; and where, in general, a responsible use of such technologies would be ensured, consistently with constitutional values. But to which extent do statements correspond to reality, i.e., are they reflected in the adopted regulatory framework? The answer may vary depending on each piece of the overall strategy, but certainly the AI Act is a major piece and its shortcomings would jeopardize the achievement of the general goal. As synthetically highlighted above, indeed, they raise serious concerns about the Act's capacity to ensure effective protection of those values, or, in other words, they raise serious doubts about its being truly (adequately) balanced and not rather paying a heavy tribute to industry and (some) governments. There is thus reason to believe that in too many and not secondary respects, either economic concerns (influenced by powerful lobbying groups) or the temptation to strengthen State control over society prevailed.

If EU's aspiration is to be "a guiding light" on fundamental rights and democracy (a model for the rest of the world) [2], then it should put the values of postwar constitutionalism over other considerations and govern technological development more effectively. Which changes would then be necessary, with specific regard to the AI Act? What has been noted in the previous sections lead us to believe that addressing

successfully the constitutional challenges of the AI era would require the strict implementation of the precautionary approach recommended by those scholars who are especially concerned by the impact of the technological revolution on constitutional values. Accordingly, European co-legislators should reduce the excessive importance given to soft law and self assessment and rather strengthen public control, namely the role of independent authorities [18], and provide effective legal remedies for the persons whose rights and dignity are affected by the use of AI systems. The safeguarding of those rights, democracy and the rule of law also requires to reduce exceptions and extend prohibitions: some systems, classified as high-risk, should instead be classified as unacceptable and consequently be banned, as their use is simply "a line no good society should cross" [2]. A regulation legitimizing the deployment of systems which are definitely harmful to human dignity only normalizes inacceptable practices and, by reducing protection when they are used on vulnerable groups, it simply allows discrimination and possibly sets the stage for a more extended use. If political representatives are not concerned by this, what remains and is needed is, firstly, referring to the Court of Justice of the EU to denounce incompliance with the Charter of Fundamental Rights and namely breaches of human dignity (this would be the case of the use of ERS in the context of migration; but other aspects of the AI Act, namely the regulation of the use by police of real-time and retrospective facial recognition in publicly accessible spaces; the categorizing of biometric data in the area of law enforcement; and predictive policy based on profiling might also be assessed as lacking the requirements of foreseeability, necessity and proportionality) [2]; secondly, the effort to increase the awareness of operators and citizens, requiring the essential contribution of educational agencies [19].

References

1. Access Now: Historic vote in the European Parliament: dangerous AI surveillance banned, but not for migrant people at the borders. https://www.accessnow.org. Accessed 14 June 2023
2. Anonymous civil servant: The AI Act—Aa breach of EU fundamental rights Charter?. EU Observer, https://euobserver.com/opinion/158050. Accessed 12 Feb 2024
3. Bennett, C.J., Lyon (eds.).: Data-driven elections: implications and challenges for democratic societies. Internet Policy Rev. **4**(8) (2019)
4. Casonato, C., Marchetti, B.: Prime osservazioni sulla proposta di regolamento dell'Unione europea in materia di intelligenza artificiale. Biolaw J. **3**, 415–437 (2021)
5. De Gregorio, G.: Digital Constitutionalism in Europe. Reframing Rights and Power in the Algorithmic Society, Cambridge University Press, Cambridge (2022)
6. Falletti, E.: Discriminazione algoritmica. Giappichelli, Torino (2022)
7. Leufer, D., Rodelli, C.: Fanny: human rights protections… with exceptions: what's (not) in the EU's AI Act deal. Access now. https://www.accessnow.org/whats-not-in-the-eu-ai-act-deal/. Accessed 15 Sep 2023
8. Mantelero, A.: Consultative committee of the convention for the protection of individuals with regard to automatic processing of personal data. Report on Artificial Intelligence. Artificial Intelligence and Data Protection: Challenges and Possible Remedies (2019)
9. Mantelero, A.: La gestione del rischio nel GDPR: limiti e sfide nel contesto dei Big Data e delle applicazioni di Artificial Intelligence. In: Mantelero A., Poletti (eds.), Regolare la tecnologia:

il Reg. UE 2016/79 e la protezione dei dati personali, pp. 289–305. Pisa University Press, Pisa (2018)
10. Mantelero, A.: La privacy all'epoca dei Big Data. In: VV. AA.: I dati personali nel diritto europeo, pp. 1181–1212. Giappichelli, Torino (2019)
11. Masala, P.: Emerging collective implications of personal data processing: challenges and responses in the European context. In: Groppi, T., Carlino, V., Milani, G. (eds.) Framing and Diagnosing Constitutional Degradation, pp. 263–272. Consulta Online, Genoa (2022)
12. Messina, D.: La proposta di regolamento europeo in materia di intelligenza artificiale: verso una "discutibile" tutela di tipo individuale di tipo consumer centric nella società dominata dal "pensiero artificiale." Media Laws **2**, 196–231 (2022)
13. Micklitz, H.-W., Pollicino, O., Reichman, A., Simoncini, A., Sartor, G., De Gregorio, G. (eds.): Constitutional Challenges in the Algorithmic Society. Cambridge University Press, Cambridge (2021)
14. Pasquale, F.: The Black Box Society. Cambridge University Press, Cambridge (2016)
15. Pollicino, O., De Gregorio, G.: Constitutional law in the algorithmic society. In: Micklitz, H.-W., Pollicino, O., Reichman, A., Simoncini, A. Sartor, G., De Gregorio, G. (eds.) Constitutional Challenges in the Algorithmic Society, pp. 3–24. Cambridge University Press, Cambridge (2021)
16. Roberts H., Floridi L.: The EU and the US: two different approaches to AI governance. Oxford Internet Institute (2021). http://www.oii.ox.ac.uk
17. Simoncini, A.: L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà. Biolaw J. **1**, 63–89 (2019)
18. Simoncini, A.: Sistema delle fonti e nuove tecnologie. Le ragioni di una ricerca di diritto costituzionale. Tra forma di Stato e forma di governo. Osservatorio sulle fonti **2**(XIV), 723–732 (2021)
19. Simoncini, A., Suweis, S.: Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale. Rivista di filosofia del diritto **1**(VIII), 87–106 (2019)
20. Tafani, D.: Do AI systems have politics? Predictive optimisation as a move away from liberalism, the rule of law and democracy. Ethics Politics (forthcoming). https://doi.org/10.5281/zenodo.10229060. Accessed 16 Feb 2024
21. Valcke, P., Clifford, D., Vilté, K.D.: Constitutional challenges in the emotional AI era. In: Micklitz, H.-W., Pollicino, O., Reichman, A., Simoncini, A. Sartor, G., De Gregorio, G. (eds.) Constitutional Challenges in the Algorithmic Society, pp. 57–77. Cambridge University Press, Cambridge (2021)
22. Veale, M., Zuiderveen Borgesius, F.: Demystifying the draft EU artificial intelligence act. Comput. Law Rev. Int. **4**(22), 97–112 (2021)
23. Zuboff, S.: The Age of Surveillance Capitalism. The Fight for a Human Future at the Frontier of Power. Profile Books, London (2019)

Chapter 19
Visual Context-Aware Person Fall Detection

Aleksander Nagaj, Zenjie Li, Dim P. Papadopoulos, and Kamal Nasrollahi

Abstract As the global population ages, the number of fall-related incidents is on the rise. Effective fall detection systems, specifically in the healthcare sector, are crucial to mitigate the risks associated with such events. This study evaluates the role of visual context, including background objects, on the accuracy of fall detection classifiers. We present a segmentation pipeline to semi-automatically separate individuals and objects in images. Well-established models like ResNet-18, EfficientNetV2-S, and Swin-Small are trained and evaluated. During training, pixel-based transformations are applied to segmented objects, and the models are then evaluated on raw images without segmentation. Our findings highlight the significant influence of visual context on fall detection. The application of Gaussian blur to the image background notably improves the performance and generalization capabilities of all models. Background objects such as beds, chairs, or wheelchairs can challenge fall detection systems, leading to false positive alarms. However, we demonstrate that object-specific contextual transformations during training effectively mitigate this challenge. Further analysis using saliency maps supports our observation that visual context is crucial in classification tasks. We create both dataset processing API and segmentation pipeline, available at https://github.com/A-NGJ/image-segmentation-cli.

A. Nagaj · D. P. Papadopoulos
Technical University of Denmark, DTU, Anker Engelunds Vej 101, 2800 Kongens Lyngby, Denmark

A. Nagaj · Z. Li (✉) · K. Nasrollahi
Milestone Systems A/S, Banemarksvej 50C, 2605 Brøndby, Denmark
e-mail: zli@milestone.dk

K. Nasrollahi
Aalborg University, Fredrik Bajers Vej 7K, 9220 Aalborg, Denmark

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_19

19.1 Introduction

According to the World Health Organization,[1] over 37 million severe falls requiring medical attention occur worldwide annually. A fall detection system can effectively mitigate the risk associated with a fall accident by allowing for a faster response time. One of the predominant strategies involves the use of sensor-based approaches, including accelerometers or other wearable sensors.[2,3,4] These devices must be worn constantly, causing discomfort over time and often requiring frequent charging. On the contrary, emerging camera-based Computer Vision (CV) systems do not require physical interaction and rely purely on captured videos or images. With the advancements in Deep Learning (DL), including robust Convolutional Neural Networks (CNN) and Visual Transformers (ViT) [2], vision-based fall detectors can be non-intrusive and applicable at scale using multi-camera setups in, i.e., hospitals or nursing homes [10].

The accuracy of a fall detector depends greatly on the available training data and operating environment. Each scene contains objects that can affect detection accuracy. Objects not central to an analysis but appearing in an image are referred as the *Visual Context* in this study. We integrate segmentation and transformation techniques to manipulate visual context and thoroughly evaluate our methods using well-established models such as ResNet-18 [5], EfficientNetV2-S [13], and Swin-Small [9]. Leveraging Segment Anything Models (SAM) [6] and saliency maps, our research provides insight into how scene elements affect detection accuracy.

The contributions of this work are multifold:

- Enhancing fall detection performance by visual context-aware augmentation during training.
- Advancing understanding of the role of visual context in fall detection, paving the way for more reliable systems.
- Creating a fall dataset with annotated visual context based on public datasets. Furthermore, we create a Python API that facilitates the usage of our or any other dataset annotated with our segmentation pipeline.

More specifically, we show that Gaussian blur applied to the background, i.e., everything except the subject, evidently improves the performance of all evaluated model architectures (Fig. 19.1).

[1] https://www.who.int/news-room/fact-sheets/detail/falls.

[2] https://www.lifeline.com/medical-alert-systems/falldetection/.

[3] https://support.apple.com/en-us/108896.

[4] https://www.medicalguardian.com/.

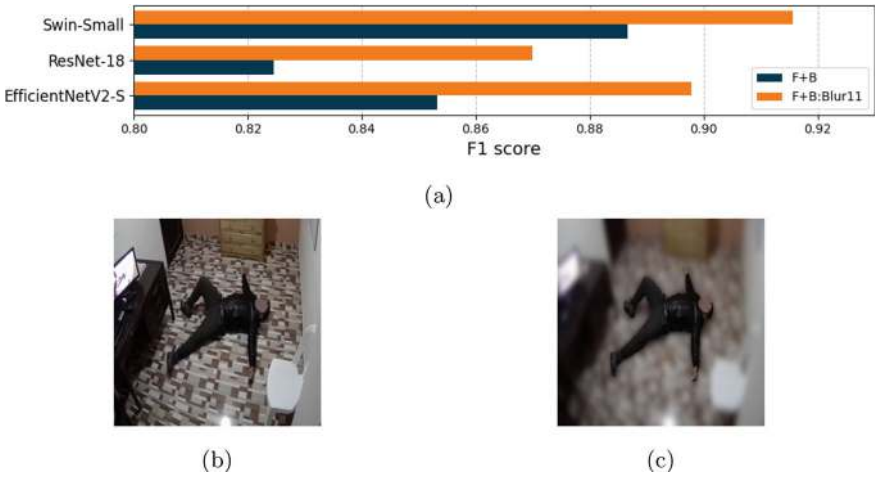


Fig. 19.1 Comparison of the performance of models trained on raw Foreground+Background (F+B) data (see Fig. 19.1b as an example), and transformed data with background smoothed using Gaussian blur with kernel size 11 (F+B:Blur11) (see Fig. 19.1c as an example). All models were tested on raw images without segmentation. The F1 score is increased for each architecture when training on transformed data. This image is sourced from the CAUCAFall dataset [3]

19.2 Related Work

19.2.1 Role of Visual Context in Computer Vision

Moyaeri et al. [11] investigated image classification models, finding that adversarial training enhances the background sensitivity of ResNets, while contrastive training diminishes foreground sensitivity. They also noted the adaptability of ViT architecture models in increasing foreground sensitivity with rising image corruption. Xiao et al. [14] analyzed the reliance of object recognition models on background signals, indicating that higher accuracy models tend to rely less on background information. Both studies, however, relied on manual segmentation methods prone to inaccuracies, focusing mainly on foreground-background separation without exploring more detailed semantic segmentation. Addressing these limitations, our research employs advanced semantic segmentation techniques for a more refined analysis, improving comprehension of complex visual contexts.

19.2.2 Advances in CV Foundation Models

Kirillov et al. [6] from Meta AI Research introduced the Segment Anything Model (SAM). It showcases exceptional generalization capabilities and flexibility in seg-

mentation prompts, making it ideal for complex downstream tasks. Liu et al. [8] developed GroundingDINO, an open-set object detector that generates bounding boxes for objects specified by a text prompt. Our research leverages both SAM and GroundingDINO to enhance object annotation in the fall dataset, demonstrating how the text-prompt guided object detection combined with advanced semantic segmentation can provide a robust labeling tool.

19.3 Dataset

There exist a few publicly available fall datasets. URFall [7] contains 70 image sequences, divided into 30 fall events and 40 activities of daily living (ADL) recorded with Microsoft Kinect cameras in an office and a private house. KULeuven [1] includes 55 fall and 17 ADL scenarios, all recorded with 5 cameras in a setting designed to resemble a nursing home. CAUCAFall [3] was captured with one camera in a home environment where each of the 10 participants performed 5 types of falls and 5 ADLs. Despite all subjects being at a relatively young age and all datasets depicting a rather limited environment, they are a valuable source of data that can be used for training and evaluation of CV-based fall detectors.

19.3.1 Our Dataset with Semantically Segmented Objects

In the development of the dataset for visual context analysis in fall detection, we extracted frames from the previously public datasets. CAUCAFall offered 7,388 fall and 12,466 non-fall images, which, after temporal downsampling and removal of ambiguous frames, yielded 1,538 fall and 1,575 non-fall images. For KULeuven, from 55 fall videos, we extracted frames at 0.5 fps for ADL and 2fps for fall events, resulting in 713 fall and 1,950 non-fall images. We chose a more frequent frame rate for falls to capture their more dynamic nature. From URFall, 53 fall and 323 non-fall images from RGB Camera 0 were distilled to 42 fall and 275 non-fall images after excluding ambiguous content. Person detection via a pre-trained YOLOx [4] model, with bounding boxes doubled in size, captured the subject and surroundings. Images were then cropped to the bounding box area. Next, using the segmentation pipeline detailed in Sect. 19.4, images were segmented, annotated, and aligned to a standardized format for consistency. Table 19.1 presents a final share of each public dataset in our dataset.

To ensure minimal information leakage between the training and testing phases, the dataset was strategically divided into training and test subsets, presented in Table 19.2. The test set contains images from URFall, a dataset with notably inferior image quality, and from three out of five KULeuven cameras, thereby introducing a varied testing scenario. The training set consists of the remaining images. KULeuven, with its multi-camera setup, facilitated the evaluation of models' capability to

Table 19.1 Share of fall and non-fall images across subsets in our merged dataset

Name	No. falls	No. non-falls	Falls + non-falls
CAUCAFall [3]	1,538	1,575	3,113
KULeuven [1]	713	1,950	2,663
URFall [7]	42	275	317
Sum	2,293	3,800	6,093

Table 19.2 Training and test set split of our fall dataset

Subset	Source	Fall	Non-fall
Training	CAUCAFall	1538	1575
Training	KULeuven cameras 3, 4, and 5	429	1174
Test	KULeuven cameras 1 and 2	284	776
Test	URFall	42	275

accurately interpret identical scenes captured from multiple angles. Additionally, the test set was intentionally imbalanced, with a lower representation of falls, to mirror real-world scenarios, where falls are less prevalent than ADL.

19.4 Methods

19.4.1 Segmentation Techniques

In the following, we describe the proposed annotation pipeline comprising GroundingDINO, SAM, and Label Studio[5] for semantic image segmentation (Fig. 19.2). The process is divided into two stages. In the first, unsupervised stage, GroundingDINO identifies key objects and produces bounding boxes for each object that are then used as input prompts for SAM which performs semantic segmentation. In this stage, we define a person, a chair, a table, a bed, a wheelchair, floor, and a walking aid (walker) as key objects. Furniture that serves a similar purpose is considered as one of the previously mentioned labels for simplification. The binary masks generated by SAM are stored in a common COCO JSON format.[6] In the second, supervised stage, masks are enhanced in Label Studio, utilizing a custom interface backed by SAM for thorough label verification and adjustment. This dual-phase approach ensures high-quality segmentation masks encoded in Run-Length Encoding (RLE) format.

[5] https://labelstud.io/.

[6] https://cocodataset.org/#formatdata.

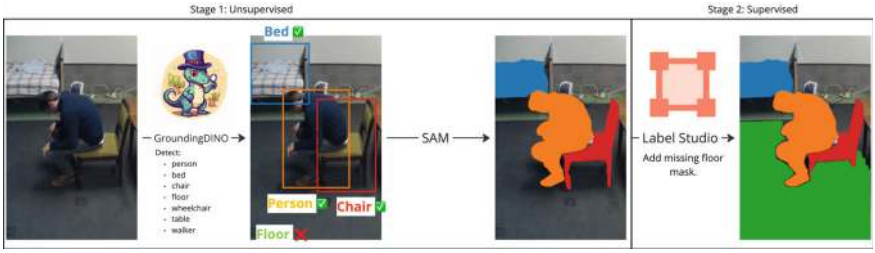


Fig. 19.2 Labeling pipeline. In the first, unsupervised stage, GroundingDINO produces bounding boxes for detected objects that are used as input prompts for SAM, resulting in segmented objects with semantic text labels. In the second, supervised stage, produced annotations are inspected and refined in Label Studio

19.4.2 Contextual Transformations

Our study focuses on modifying contextual information to gain meaningful insights into how visual context affects fall detectors. We focus on a data-centric approach, applying pixel-based transformation methods to the labeled dataset before model training. Transformations include: **solid color** (to assess model reliance on background information), **Gaussian blur** (to test model sensitivity to detail clarity), and **grayscale** (to evaluate the importance of color information). Figure 19.3 presents examples of transformations applied to one or more key objects. First, a transformation is applied to the entire image, and then a binary mask representing one or more key objects is applied, leaving the remaining parts of the image intact. Additionally, we apply augmentation methods in the following order: resize to 256×256, random perspective with distortion scale 0.4 (training only), random horizontal flip (training only), transform to PyTorch tensor, normalize. Those augmentation methods are applied *after* contextual transformations except Gaussian blur, which is applied either before or after resizing the image to a fixed resolution. In this way, we have devised two scenarios: one involving controlled smoothing (image resizing before

(a) Background (person inverse): Gaussian blur.

(b) Person (foreground): solid black.

(c) Background: grayscale.

Fig. 19.3 Pixel-based transformation applied to different areas of an image

applying image blur) and another featuring a seeded kernel size (image resizing after applying image blur).

In our evaluation, we utilize the F1 score as the primary metric. This metric is particularly suitable for assessing imbalanced datasets, which is typical in fall detection tasks where actual fall events are infrequent compared to normal activities.

19.5 Experiments

Each model was trained with the following hyperparameters: batch size 32, learning rate 0.001, momentum 0.9, 50 epochs, and early stopping on validation loss with patience 5. To mitigate the risk of having nearly identical frames in both validation and training sets, images were initially grouped into sets of five consecutive frames. These groups were then randomly split, maintaining a 10% allocation for the validation set.

19.5.1 Nomenclature

In this context, "Foreground" (F) refers to the main object of interest, typically a person, while "Background" (B) encompasses everything outside the foreground, essentially the inverse of the foreground. The nomenclature for test scenarios is as follows:

$$F : [\text{transform}] + B : [\text{transform}]$$

Square brackets [] denote optional parameters. Transformation applies to one or key objects or their inverse.

19.5.2 Effect of Gaussian Blur Transformation

To implement a Gaussian blur transformation, there is a need to specify a kernel size. We train each model on images with background transformed with Gaussian blur with increasing kernel size F+B:Blur$x$, adhering to condition $x = 2p + 1$, $p \in \mathcal{N}$, $3 \leq x \leq 31$. We employ two Gaussian blur strategies as described in Sect. 19.4.2. Models were evaluated on the test set F+B, without contextual transformations applied, unless otherwise stated.

As shown in Fig. 19.4, a seeded kernel, which allows for some degree of randomness in smoothing strength, results in significantly more robust results than a fixed kernel. Moreover, a modest kernel size often performs best. Without jeopardizing the generality of the study, we proceeded with a seeded kernel Gaussian blur with a kernel size of 11 in further experiments.

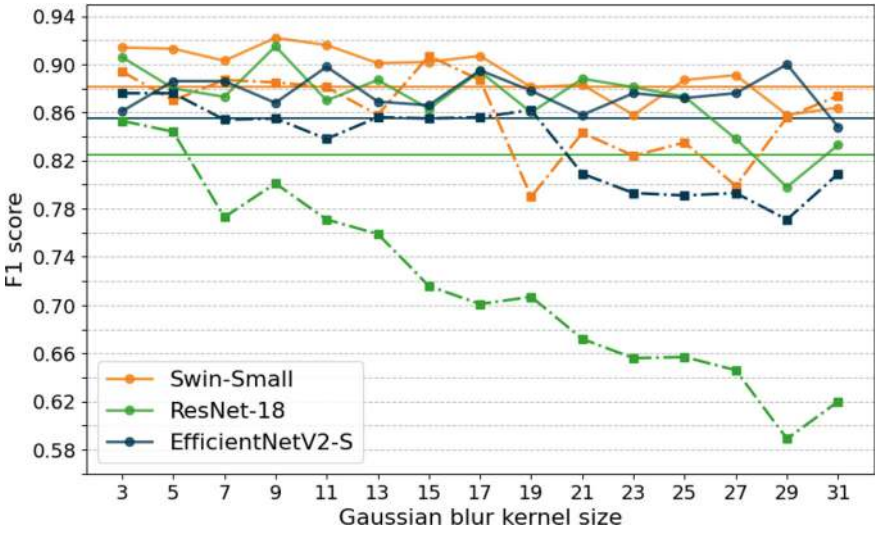


Fig. 19.4 F1 scores over Gaussian blur kernel size for different model architectures. Horizontal lines represent the F1 score for baseline F+B model (trained without contextual transformations). Solid lines with dot markers represent models trained with seeded kernel (blur before resize). Dashdot lines with square markers represent models trained with a fixed kernel (blur after resize)

19.5.3 Understanding the Influence of Visual Context

To understand the impact of visual context on fall detector performance, we trained each model on transformations (see Sect. 19.4.2), applied to the foreground or background. Table 19.3 presents evaluation results on the F+B test set. We observe marked robustness to blur transformations, where training on images with blurred backgrounds boosts performance on raw F+B data, suggesting that smoothing supports more generalizable feature learning. Conversely, drastic alternations, like a solid black transformation, degrade performance and lead to overfitting. CNNs are more reliant on color information than transformers as grayscale transformation causes a significant drop in the performance of both ResNet-18 and EfficientNetV2-S. This analysis highlights the dominance of foreground information over the background but stresses the importance of both for optimal visual data interpretation, showcasing their complex interplay in classification tasks.

Table 19.3 F1 Scores for the F+B test scenario. **Bold** indicates the best test for a model. The F1 scores in parentheses are evaluations on the test set with the same contextual transformations as in model training

Test \ Train	F+B	F+B:Blur11	F+B:SolidBlack	F+B:Grayscale	F:Blur11+B	F:SolidBlack+B	F:Grayscale+B
F+B ResNet-18	0.825	**0.87** (0.919)	0.551 (0.912)	0.724 (0.871)	0.807 (0.898)	0.638 (0.867)	0.722 (0.854)
F+B EfficientNetV2-S	0.853	**0.898** (0.83)	0.471 (0.872)	0.777 (0.87)	0.825 (0.842)	0.566 (0.824)	0.787 (0.841)
F+B Swin-Small	0.887	0.916 (0.916)	0.732 (0.839)	0.913 (0.895)	**0.929** (0.873)	0.743 (0.872)	0.926 (0.904)

19.5.4 Significance of Specific Objects

We identify key objects as those often found in misclassified images, suggesting a model's difficulty in correctly classifying a fall when these objects are present. We assessed F+B and F+B:Blur11 models on test subsets with key objects for each architecture. Across all models, there was a notable response to the presence of key objects (Fig. 19.5). Wheelchairs posed a significant challenge for ResNet-18 and Swin-Small, whilst beds were particularly problematic for ResNet-18 and EfficientNetV2-S. People interacting with beds or wheelchairs may assume poses similar to a fall, potentially confusing a fall detector. Smoothing the background has proven effective in reducing such impact, suggesting that a less detailed background enables the fall detector to better focus on the primary subject, thereby improving classification accuracy.

19.5.5 Qualitative Insights from Saliency Maps

GradCAM [12] is commonly utilized to visually explain CNN-based models by producing a heatmap using gradients from the selected convolutional layer, typically the final one. We explored how attention changes when training the model on a smoothed background. Due to implementation limitations, saliency maps in our study are created only for CNNs. We used default GradCAM parameters and applied them to the last convolutional layer. As shown in Fig. 19.6, the model tends to shift its attention toward the person when trained on the transformed visual context, indicating improved subject distinction through background blurring.

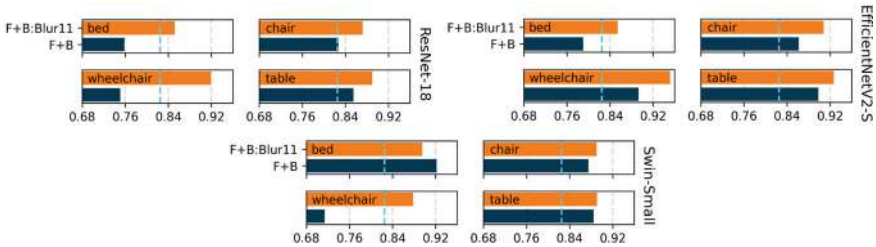


Fig. 19.5 F1 scores for subsets of the training set containing the specific key object for each analyzed model architecture. The light blue dashed line shows the F1 score of an F+B trained model for each architecture, tested on the entire dataset

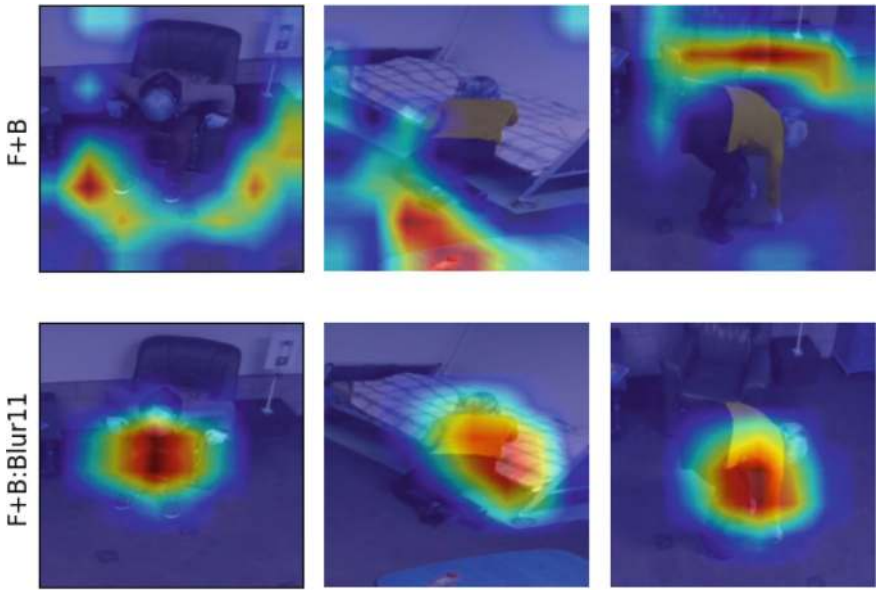


Fig. 19.6 Saliency map examples from GradCAM for the last convolutional layer of EfficientNetV2-S. Top row: F+B models, bottom row: F+B:Blur11 models. The heatmap scale ranges from blue (lowest) to red (highest) influence

19.6 Conclusions

Visual context is highly relevant for deep learning-based fall detection classifiers. Objects such as wheelchairs and beds tend to have a strong influence on their performance. Properly applied context-aware transformations to the training data benefit the accuracy and generalization capabilities of fall detectors, leading to more reliable systems. Further effort is required to comprehend the extent and working mechanism of the impact of visual context and its applicability to other computer vision tasks. Additionally, expanding datasets to encompass diverse conditions and addressing ethical concerns are essential for advancing fall detection technologies.

References

1. Baldewijns, G., Debard, G., Mertes, G., Vanrumste, B., Croonenborghs, T.: Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. Healthc. Technol. Lett. **3**(1), 6–11 (2016). https://doi.org/10.1049/htl.2015.0047
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale (2021)

3. Eraso, J.C., Muñoz, E., Muñoz, M., Pinto, J.: Dataset Caucafall (2022). https://doi.org/10.17632/7w7fccy7ky.4
4. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding Yolo Series in 2021 (2021). arXiv:2107.08430
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2015)
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything (2023)
7. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
8. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding Dino: Marrying Dino with Grounded Pre-training for Open-Set Object Detection (2023). arXiv:2303.05499
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10,012–10,022 (2021)
10. Madsen, K., Li, Z., Lauze, F., Nasrollahi, K.: Person fall detection using weakly supervised methods. In: WACV Real World Surveillance Workshop (2024)
11. Moayeri, M., Pope, P., Balaji, Y., Feizi, S.: A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds, Backgrounds, and Visual Attributes (2022)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
13. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
14. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or Signal: The Role of Image Backgrounds in Object Recognition (2020)

Chapter 20
Addressing Racial Bias in AI: Towards a More Equitable Future

Elisabetta Ferrara

Abstract The widespread diffusion of Artificial Intelligence (AI) technologies indicates a transformative tension that raises concerns about implicit racial bias and reignites debates around the concept of race to justify a presumed superiority of one part of humanity over another, and its repercussions in the scientific field. This dispute highlights the need to explore the connections between institutional and socio-cultural dimensions, projecting the discussion onto a global sphere that encourages reflection on human dimensions and the concept of technological determinism. This review, through a methodical analysis of recent literature ranging from empirical studies to theoretical discussions, aimed to examine the intricate interplay between AI/ML and the principle of non-discrimination, with a specific focus on the innate racial biases within algorithms in AI/ML technologies, their extensive implications, and subsequent outcomes.

20.1 Introduction

In the contemporary digital era, data-intensive Artificial Intelligence (AI) systems are increasingly scrutinized for their potential to erode critical aspects of ethical principles, particularly regarding non-discrimination [1]. Characterized by their reliance on extensive datasets to train algorithms, these AI systems have the capability to process, analyze, and make decisions based on the accumulated data on an unparalleled scale [2]. This evolving scenario significantly raises concerns about the control individuals maintain over their personal information, a cornerstone of human rights. In our rapidly digitizing world, this emerging dynamic in different spheres presents both monumental opportunities and significant ethical questions, especially concerning the principle of non-discrimination. AI's decision-making capabilities, powered by sophisticated algorithms, have become instrumental in influencing various aspects of

E. Ferrara (✉)
Telematic University "Leonardo Da Vinci" (UNIDAV), Torrevecchia Teatina, Chieti, Italy
e-mail: elisabetta.ferrara@unidav.it

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_20

human life, from the content we interact with on social media platforms to the determination of eligibility for state assistance [3, 4]. The intent behind deploying AI technologies generally revolves around maximizing operational efficiency and scalability of systems. However, this technological leap brings forth the critical issue of potential bias within AI algorithms and their discriminatory implications, particularly against vulnerable groups such as women, immigrants, and racial minorities [4, 5]. The EU Agency for Fundamental Rights has underscored that the deployment of AI can intersect significantly with numerous fundamental rights [6]. While AI has the potential to serve beneficial purposes, its improper application can infringe upon privacy rights and lead to decisions marred by discrimination, thereby profoundly impacting individuals' lives. An advisory example from the Netherlands in 2020 highlighted how biased algorithmic operations by tax authorities unjustly targeted approximately 26,000 parents over childcare benefit fraud, disproportionately affecting those with an immigrant background and imposing severe financial and psychological strains upon the families involved. Despite increasing awareness of algorithmic biases, the field still grapples with a scarcity of concrete evidence that critically evaluates the functioning and outcomes of these algorithms in real-world contexts. There's a pressing need for rigorous, evidence-based assessments to bridge this knowledge gap and thoroughly understand how AI might contravene fundamental rights [2, 7]. The European Commission's proposal for an AI [1, 3, 8], aiming to safeguard fundamental rights within the AI context, sets a legislative focus on this issue. This report was aimed to enlighten policymakers, human rights practitioners, and the broader public on the intricacies of bias within AI, advocating for AI applications that through respect and uphold fundamental rights. Specifically, it investigated into the development of bias feedback loops in predictive policing and the challenge of detecting bias in algorithms designed to identify offensive speech, presenting an in-depth analysis to detect and mitigate forms of bias that could lead to discrimination, thereby aligning AI algorithms with the core principles of non-discrimination and fundamental rights protection [9]. This review aimed to explore how racial bias, a specific yet pervasive form of algorithmic bias, intersects with AI technologies and the consequent challenges it poses to the principle of non-discrimination. By focusing on concrete manifestations of racial bias in medical AI applications, this review elucidates the intricate ways biases are embedded within algorithms, complicating the assessment and mitigation of potential discrimination. These inquiries are vital in navigating the complexities of ensuring AI serves as a conduit for equity and justice, rather than an instrument of discrimination. By focusing on the medical area, we explored how biases manifest within algorithms at various stages, complicating the assessment of potential discrimination.

20.2 AI and Racial Bias

Racial bias in Artificial Intelligence (AI) represents one of the most critical and pressing issues to address in contemporary technological evolution. AI and Machine Learning (ML), a cornerstone of AI, while offering revolutionary capabilities in areas ranging from facial recognition to health care, reflect and risk amplifying the biases and inequalities present in society [2, 7, 10]. This occurs primarily because AI algorithms are trained using datasets collected from human societies, which can contain historical and social biases. Racial bias in AI manifests in various ways, adversely affecting people's lives. For instance, in the area of facial recognition, studies have shown that some systems struggle more with correctly recognizing faces of non-White individuals due to training datasets that predominantly include White faces. Implicit and unconscious biases manifest in AI and in ML, applications through various mechanisms, often leading to the perpetuation of racial discrimination [11]. The Kirwan Center's definition of implicit bias clarifies how attitudes or stereotypes, operating below the level of conscious awareness, can affect decisions and actions. When these biases are transferred to technological domains, they contribute to the challenges of racial bias within ML and AI. The integration of implicit biases into ML and AI systems primarily occurs through the data and algorithms that underpin these technologies [5, 12]. Training data for ML algorithms, reflecting historical and societal biases, can cause AI systems to inadvertently replicate and even amplify discriminatory practices [13]. The absence of diverse perspectives in tech development teams can lead to "inattentional blindness," where developers might overlook biases in AI systems due to their cultural or experiential homogeneity. This phenomenon was notably highlighted when AI algorithms, such as those used in facial recognition technologies and language processing applications, displayed biased results, misidentifying African Americans or associating them with negative stereotypes [14]. Within the digital era, research has highlighted inaccuracies and societal stereotypes embedded within these predictive algorithms, which disproportionately depict African Americans as more prone to engaging in violent crimes in comparison to their White counterparts [15]. The discussion extends to the racial neutrality of predictive sentencing models like the COMPAS algorithm that quantifies a defendant's likelihood of future criminal activity with risk scores. Findings suggest that individuals assigned higher risk scores, which ostensibly predict a higher likelihood of recidivism, face greater pre-trial detentions, underscoring a systematic bias that disparately impacted African Americans [5]. These biases create erroneous associations between genders and professional fields, as well as between racially coded names and negative or positive attributes, thus perpetuating stereotypes [16]. Online data collection, encompassing interactions with websites, social media, and e-commerce platforms, fuels the vast reserves of big data, significantly impacting fields like science, health care, and governance. This digital profusion, however, harbors potential for misuse, particularly in the misapplication of big data analytics which can lead to privacy invasions and discriminatory outcomes based on users' online activities. The Federal Trade Commission has highlighted instances

where such analytics have resulted in unjust profiling, impacting individuals' access to credit, employment, and educational opportunities based on biased data interpretations. Efforts to counteract such biases, while present, face limitations inherent in the current regulatory frameworks. Notably, laws like the Civil Rights Act and the Fair Housing Act address explicit discrimination but fall short in mitigating the more subtle, implicit biases entrenched in algorithmic design. This setting of big data utilization highlights the complex interplay between technological advancement and its societal implications, necessitating a vigilant approach to ensuring technology serves equitable ends.

20.3 The Algorithm Prejudice in Racial Bias

An equitable algorithm is characterized by its ability to accord identical scores or classifications to patients with comparable fundamental needs. In the healthcare sector, algorithms find application in an array of functions including diagnostics, therapeutic decisions, prognosis evaluation, risk categorization, prioritization, and the distribution of resources [2, 4]. Instances of algorithmic prejudice included a model that, by incorporating race in its calculations for kidney functionality, yielded inflated figures for Black patients over White patients, thereby hindering timely organ transplant referrals for the former group. Additionally, a commercial algorithm designed for risk assessment in chronic disease management inadvertently stipulated more severe illness criteria for Black individuals than for White individuals to access respective programs. Algorithms with potential biases have been formulated across various medical fields such as cardiology, nephrology, obstetrics, oncology, affecting the accessibility to or qualification for medical interventions and the allocation of resources [8]. Interestingly, a study conducted by Shanklin et al. [17], exploring the for racial bias within ML-driven appointment scheduling algorithms highlighted how conventional scheduling objectives, prioritizing solely the minimization of patient waiting times and provider idle/overtime periods, may disproportionately affect patients with higher no-show probabilities, potentially associated with specific racial groups. This has the potential to create unintentional racial disparities in wait times and negatively impact the service experience for these patients. Through rigorous analysis of both simulated and real-world data from a medical clinic, the study revealed that existing scheduling systems can lead to significant racial disparities, with Black patients demonstrably waiting up to 30.

20.3.1 Ethical and Empirical Challenges of Algorithmic Equality in Health Care

The combination of AI and ML, which is a fundamental aspect of AI, has shown notable ethical issues, namely, related to the continuation of biases inherent in health data. Multiple strategies have been developed in order to achieve algorithmic fairness, with the goal of constructing unbiased models. These models aim to produce neutral predictions by addressing bias against protected identities, such as race or gender, through various techniques [11, 17] These methods include excluding sensitive variables (which can unintentionally increase discrimination), equalizing error rates among different groups, ensuring that outcomes are independent of identities after risk assessment, and establishing a mathematical balance between benefits and disadvantages across different demographics. Although there is a strong need to incorporate ethical issues into algorithm design and a growing business trend towards impartiality solutions, this method has inherent drawbacks [4]. An important omission in algorithmic fairness is its insufficient consideration of the complex causal connections between biological, environmental, and social factors that contribute to health disparities among marginalized groups. Although it is undeniable that these social determinants have an impact on health outcomes, there is still a lack of clear understanding of the specific effects, particularly when considering overlapping identities [2, 7]. Moreover, the distinction that these models provide between identities does not necessarily result in unfairness. The implementation of algorithmic equality solutions in health care is filled with uncertainty regarding the connection between protected identities and health outcomes, resulting in practical difficulties. An example of this is the lack of representation of women, especially women of color, in cardiac research [18]. This demonstrates how the exclusion of some groups from data can result in unequal progress in the effectiveness of treatments for different genders. Several approaches to achieving fairness unwittingly perpetuate discriminatory practices by seeking to represent minority groups using the traits of the majority. These approaches overlook the underlying variety and rely on a simplified universal standard. Moreover, when impartiality-corrected models make predictions about treatment outcomes using data from normative groups, there is a possibility that these models may not accurately forecast the reactions of individual patients whose answers differ from the standardized projections. This might reduce the usefulness of these models in a clinical setting and potentially hide important interventions. The study conducted by Giovanola et al. [19], presents new insights on the concept of neutrality in AI, particularly in the context of Healthcare Machine Learning Algorithms (HMLA). This study identified a significant lack of understanding by suggesting that fairness in AI should not be limited to just eliminating bias or following a distributive justice paradigm, but should also include a wider ethical aspect. The authors have developed a strong and systematic approach to understanding fairness by incorporating ideas from moral philosophy. They argue that fairness is an intrinsic ethical goal that encompasses principles like fair equality of opportunity, the difference principle, and the equal right of justification. These

concepts are considered crucial for driving the ethical design and implementation of HMLA, ensuring that these technologies uphold both individual and society ethical standards beyond mere bias moderation.

20.3.2 The Impact of Racial Bias in AI on Healthcare Diagnostics

Bias and discrimination within AI frameworks have been scrutinized across multiple sectors, including an array of healthcare deployments ranging from melanoma detection to mortality forecasts and anticipatory models for healthcare engagement [6, 17]. From the identification of melanoma to predicting mortality and developing models for proactive healthcare involvement. These studies reveal a division in the effectiveness of AI based on self-reported racial categories in several therapeutic tasks [5]. Research has illuminated the disparity in medical AI systems' performance contingent upon race; for instance, Gichoya et al. [20], documented discernible disparities in the precision of automated diagnostics for chest X-rays across diverse racial and demographic segments, despite the models' exclusive interaction with the radiographic images. Critically, the implementation of such prototypes would result in a disproportionate misclassification of Black and female patients as healthy compared to their White and male counterparts. Furthermore, these racial disparities cannot solely be attributed to the under-representation of these demographic groups in the training datasets, as no statistically significant correlation has been established between membership in these groups and the observed racial disparities. Complementary research has revealed the capacity of AI algorithms to discern various demographic facets of patients [11]. One investigation uncovered an AI model's ability to predict gender and differentiate between adult and pediatric patients from chest X-rays, whereas other studies have demonstrated reasonable accuracies in estimating patients' chronological ages from a range of imaging modalities. Within the field of ophthalmology, retinal imaging has been utilized to forecast gender, age, and markers of cardiovascular well-being, such as hypertension and smoking habits. These insights, which underscore a strong correlation between demographic variables known to influence disease trajectories (e.g., age, gender, and racial identity) and the characteristics identified in medical imagery, might inadvertently introduce bias into model outcomes, reflecting over a century of clinical and epidemiological research highlighting the crucial role of covariates and potential confounding factors [13].

20.3.3 Ethical and Equity Dimension of Racial Prediction in Medical Imaging AI

This technological progress exposes notable ethical and equality concerns, particularly about the hidden implicit biases that these AI systems may possess. The ambiguity around AI reveals its dual nature: although technology brings significant progress in medical diagnosis, it also poses the potential of reinforcing pre-existing racial biases. The current research [1–5, 12, 19, 20], thoroughly investigates the advanced capability of deep learning models to accurately determine patient ethnicity using just medical images. These studies challenge the prevailing belief about the limitations of both human specialists and AI algorithms in this regard. These findings demonstrate the effectiveness of advanced models in consistently predicting ethnicity in various clinical settings, imaging techniques, and patient demographics. By using both public and private datasets, it has been demonstrated that these models are able to overcome the differences in imaging methods specific to different racial identities and revealed that disease distribution and patient body habitus in the analyzed datasets do not determine racial group affiliation. This suggests that deep learning algorithms can infer ethnicity without relying solely on these traits. These insights provide clarity to the discussion surrounding the reduction of differences in model performance. They thoroughly analyze the possibility of making AI models "colorblind" by selectively eliminating sensitive features. However, this approach may not be feasible in medical imaging because it is difficult to separate race-related information. This discourse offers a critical viewpoint on the issue of racial biases in AI applications in health care. It suggests that instead of solely relying on technical solutions to address bias, a more sustainable approach would involve intentionally designing models that aim to achieve equal outcomes for different racial groups. This proposal aligns with the emerging regulatory landscape, which currently lacks effective mechanisms to prevent unintentional racial recognition or mitigate potential harms. Recent progress in the field of deep learning, specifically its application to chest X-rays, is extensively documented in literature [20]. Despite this, the mere presence of substantial datasets does not inherently guarantee models' effective generalization. Illustrating this challenge, researchers trained Convolutional Neural Networks (CNNs) using data from two different institutions only to discover the models struggled to generalize when evaluated on data from a third institution. This limitation in generalization may be partially attributed to the CNNs leveraging features not intrinsically linked to the intended diagnoses, such as irrelevant correlations that might exist in the training data, for instance, patient gender. Such occurrences underscore broader concerns about bias in ML a subject scrutinized across multiple facets within healthcare contexts by several researchers [8, 12], Among notable findings, Obermeyer et al. [21] revealed racial bias within an algorithm widely used across the U.S. healthcare system, demonstrating discrepancies in illness severity between Black and White patients assigned identical risk levels. Contrary to these studies, focused pivots to biases emanating from incidental correlations within training datasets that

do not consistently apply to external test sets, leading to CNNs potentially misapplying these correlations during the learning process but faltering upon application to unseen test data. This phenomenon, referred to as "shortcut" learning by CNNs, transcends health care, with substantial research devoted to understanding and mitigating biases and shortcuts across various applications. Addressing these concerns, especially in diagnostic tasks leveraging chest X-rays, necessitates exploring mechanisms to inhibit CNNs from exploiting such "shortcuts". Traditional data resampling to align the training and test sets, might improve generalization but lacks scalability, particularly in diverse institutional settings or without access to ample data for resampling. Transfer learning emerges as an alternative, leveraging pretrained models to initialize weights for related target tasks, in transferring ImageNet-trained features for interstitial lung disease classification. In parallel, multitask learning proposes simultaneous knowledge acquisition across different tasks, a method that has shown promise in various domains including health care. This approach could offer a pathway to mitigating bias by integrating information between related source and target tasks.

20.4 Conclusions

The duality uncovered necessitates immediate policy action and the implementation of ethical guidelines to ensure that AI technologies contribute to advancing the principles of justice and equality in health care, thereby reducing racial disparities in care.

References

1. Noseworthy, P.-A., Attia, Z.-I., Brewer, L.-C., Hayes, S.-N., Yao, X., Kapa, S., Friedman, P.-A., Lopez-Jimenez, F.: Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circ. Arihythm. Electrophysiol.
2. Naudts, L.: How machine learning generates unfair inequalities and how data protection instruments may help in mitigating them. In: Leenes, R., van Brakel, R., Gutwirth, S., De Hert, P. (eds.) Data Protection and Privacy: the Internet of Bodies, pp. 71–92. Hart Publishing (2019)
3. Intahchomphoo, C., Gundersen, O.-E.: Artificial intelligence and race: a systematic review. Leg. Inf. Manag. **20**(2), 74–84 (2020)
4. Lee, N.T.: Detecting racial bias in algorithms and machine learning. J. Inf. Commun. Ethics Soc. **16**(3), 252–260 (2018)
5. Huang, J., Galal, G., Etemadi, M., Vaidyanathan, M.: Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. JMIR Med. Inform. **10**(5), e36388 (2022)
6. European Commission.: Coordinated plan on artificial intelligence. Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the Regions, COM (2018) 795 final, Brussels (2018)
7. European Commission.: Communication from The Commission to The European Parliament. The European council, the council, the European economic and social committee and the

committee of the regions. Artificial intelligence for Europe. SWD (2018) 137 final. Brussels, 25.4.2018 COM (2018) 237 final (2018)
8. Leavy, S., O'Sullivan, B., Siapera, E.: Data, power and bias in artificial intelligence (2020). arXiv:2008.07341
9. Benthall, S., Haynes, B.D.: Racial categories in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 289–298 (2019)
10. Charter of Fundamental Rights of the European Union.: (2016)
11. European Commission. White Paper on artificial intelligence—A European approach to excellence and trust. Brussels, 19.2.2020 COM(2020) 65 final (2020)
12. Angileri, J., Brown, M., DiPalma, J., Ma, Z., Dancy, C.L.: Ethical considerations of facial classification: reducing racial bias in AI (2019). Accessed 21 Feb. 2020
13. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med. **169**(12), 866–872 (2018)
14. Sengupta, K., Srivastava, P.R.: Causal effect of racial bias in data and machine learning algorithms on user persuasiveness & discriminatory decision making: an empirical study (2022). arXiv:2202.00471
15. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
16. Shanklin, R., Samorani, M., Harris, S., Santoro, M.A.: Ethical redress of racial inequities in AI: lessons from decoupling machine learning from optimization in medical appointment scheduling. Philos. Technol. **35**(4), 96 (2022)
17. Kundi, B., El Morr, C., Gorman, R., Dua, E.: Artificial intelligence and bias: a scoping review. AI Soc. 199–215 (2023)
18. Giovanola, B., Tiribelli, S.: Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Soc. **38**(2), 549–563 (2023)
19. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C. et al.: AI recognition of patient race in medical imaging: a modelling study. Lancet Digit. Health **4**(6), e406–e414 (2022)
20. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019); Raymond, A.H., Shackelford, S.J.: Technology, ethics, and access to justice: should an algorithm be deciding your case. Mich. J. Int. L. **35**, 485 (2013)
21. Gentzel, M.: Biased face recognition technology used by government: a problem for liberal democracy. Philos. Technol. **34**, 1639–1663 (2021)

Part III
Reasoning-Based Intelligent Applied Systems

Chapter 21
A Study on Outlier Correction Techniques Using Multi-agent Techniques for the Accurate Predictions of Human Mobility

P. P. G. Dinesh Asanka, Masakazu Takahashi, and Chathura Rajapakshe

Abstract This research is part of the design and implementation of an AutoML platform for time series forecasting. Having discussed missing value imputation and outlier detection in other research papers, this research is focused on outlier correction to forecast better in time series. This research has used human mobility data as a sample that was collected in Hiroshima, Japan, over a year. Master–slave multi-agent design pattern was used with an outlier coordinator agent that direct other sub-agents to identify the best regression technique and correct the outlier data points. Extensible regression agents are used to identify the best regression technique using seasonal data and evaluated with the harmonic mean of regression evaluation parameters. Outlier correction was done iteratively to find the optimum outlier corrections, as fixing all the outlier data points is impossible. In this research, different levels of outlier corrections were achieved for the ten selected locations in Hiroshima, Japan. These decisions are integrated with classification techniques as an emerging knowledge capability in the multi-agent.

21.1 Introduction

In the modern world, a lot of domains will benefit from forecasting human mobility. Forecasting human mobility benefits areas such as security, marketing, and health, among other sectors. This research is part of a major research work that introduced knowledge-driven multi-agent technology for automated machine learning in time series analysis in the context of human mobility. Time series forecasting

P. P. G. D. Asanka (✉) · C. Rajapakshe
Department of Industrial Management, University of Kelaniya, Kelaniya, Sri Lanka
e-mail: dasanka@kln.ac.lk

M. Takahashi
University of Yamaguchi, Yamaguchi, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_21

and Kalman filters are the common techniques that are employed to forecast human mobility. Considering the complexities of human mobility, such as missing values, outliers, seasonality, and volume of the data, it is decided to employ time series for human mobility forecasting over Kalman Filter [12]. Existing human mobility data should be converted to time series continuous data to forecast human mobility so that time series techniques can be employed. Missing value imputation, outlier detection and correction, and normalization are common pre-processing techniques for time series forecasting. The previous publications discussed missing value imputation and outlier detection pre-processing techniques in detail. Even though outlier detection will provide essential findings, outliers should be corrected for better forecasting. This research paper focuses on producing a method to correct outliers in the time series data.

Outlier correction is considered partially observable as there is noisy data. Further, the environment is uncertain and dynamic, as fixing outliers may result in additional outliers. Due to these complexities, it is necessary to implement multi-agent architecture.

This research paper is organized into the literature review, where major areas of the research are discussed, and the methodology section discusses the multi-agent design patterns that are used. The implementation section has discussed the multi-agent implementation and evaluation of techniques.

21.2 Literature Review

By considering the research areas, it was decided to carry out a literature review in the areas of multi-agent technologies, time series techniques, outlier correction techniques, and knowledge-emerging techniques. Even though AutoML has emerged in many areas like classification, regression, time series, rankings [17], vision [1], etc. A lot of industry tools, such as Microsoft Azure [2] Google, and Mathworks, have introduced AutoML to their tools stack. In contrast, new AutoML frameworks such as AutoML frameworks such as Microsoft AutoML [6], EvalML [8], AutoML Bench [7], AutoWeka [10], TPOT [16], and MindsDB [13] have emerged. Even though there are various platforms to forecast time series data such as AzureML, it was noticed that many of the platforms are limited to selecting the best forecasting algorithm by evaluating a handful of evaluation techniques.

As time series forecasting needs many pre-processing tasks such as stationery testing, missing value imputation [9, 11, 14], normalization [15], outlier detection, and outlier correction, those tasks should be considered as the part of the AutoML platform.

Multi-agent modeling is proposed with knowledge-emerging and decision-making properties. Out of existing multi-agent design patterns, the master–slave or message space agent method [3] is proposed when complex and extensible sub-agents are required.

This research paper is part of a significant research that proposes AutoML platform for time series forecasting. In this research, missing value imputation [4] and outlier detection [5] are already completed and this research paper is focused on outlier correction.

21.3 Multi-agent Architecture

This research has proposed a comprehensive multi-agent architecture for time series forecasting, including data procession missing value imputation data normalization, outlier detection, outlier correction, and forecasting, as shown in Fig. 21.1. This research has used human mobility dataset at Hirishima between 2019 December and 2020 November. Considering the COVID-19 pandemic, there were considerable outliers in the dataset. This one million dataset was converted to time series data for seven locations: Hiroshima Castle, Hiroshima Botanical Garden, Hiroshima Station, Ballpark, Hiroshima Prefectural Office, Hiroshima MOCA, Bomb Site, Hiroshima City Hall, and Hachobori Hiroshima.

In the proposed architecture, each multi-agent has its own knowledge base to help its decision-making process. As of today, missing value imputation and outlier

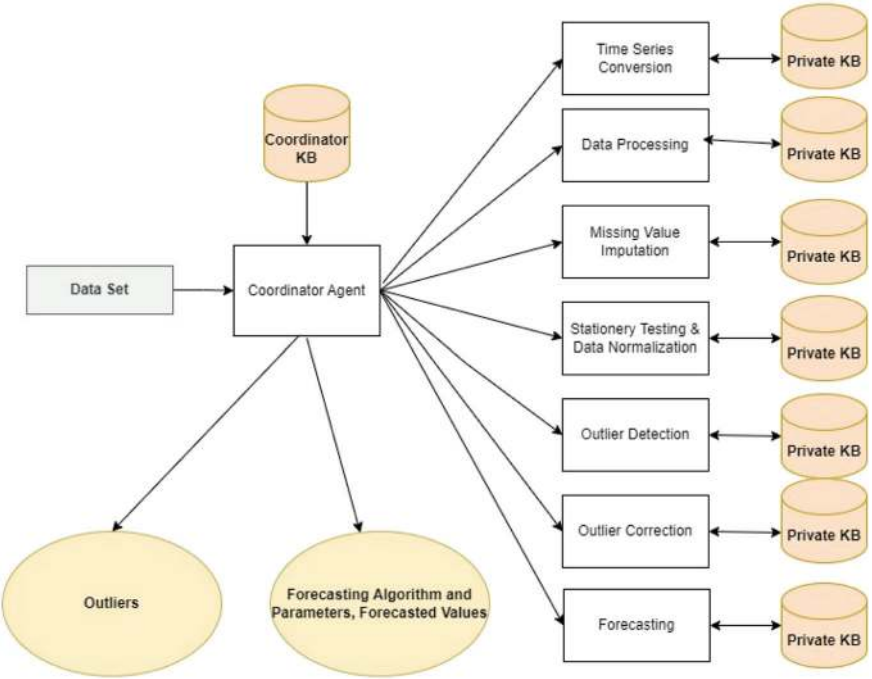


Fig. 21.1 Proposed multi-agent architecture for time series forecasting

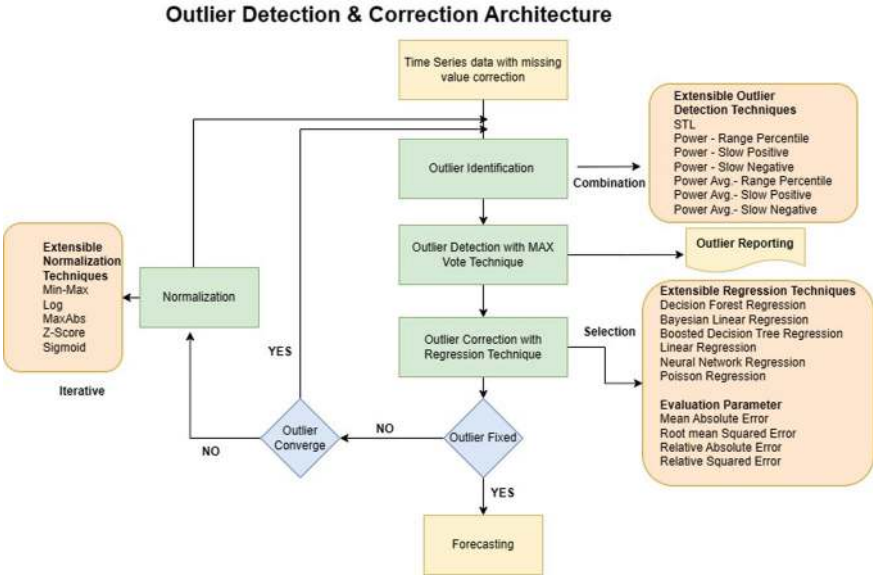


Fig. 21.2 Flow diagram for outlier detection and correction

detections have been completed. Even though the outlier detection itself is a valuable outcome, correcting the outliers is important for better forecasting time series.

In the previous research, AutoML multi-agent platform was proposed and this research is the extension of the previous research, as shown in Fig. 21.2.

As indicated in the Fig. 21.2. Seven different techniques were used to detect the outliers and maximum vote techniques were used to report the outliers. Those outlier points are corrected by the regression technique. To apply a regression techniques, different types of seasonal factors such as location, day, hour, month, day of the week, visits at the previous hour, visits at the next hour, visits at the next week, visits at the previous week, visits at the next month, and visits at the previous month were considered. Unlike the outlier detection technique, to select the best regression technique different techniques were employed. Decision forest regression, Bayesian linear regression, boosted decision tree regression, linear regression, neural network regression, and Poisson regression techniques were evaluated and the best technique was selected for the technique that has a minimum harmonic mean of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Relative Squared Error (RSE). The selected regression technique is used for outlier correction.

When a few outlier points are corrected, there can be silent outliers that were not identified before but will be outliers now. Even though this can be avoided if the prediction is done using the noise data, the prediction of noise data is not a recommended technique in time series. This means outlier correction cannot be done only with one iteration but needs to follow multiple iterations until the optimum

number of outliers is corrected. Ideally, with each iteration, the number of outliers should be reduced. However, there can be instances where the number of outliers will not be converged. In that situation, normalization techniques should be used that will be considered as a future work in this research.

In this research, multi-agent techniques are utilized as a solution for the AutoML platform for Time Series Forecasting. One of the main reasons to introduce multi-agent is that it can work in minor agents in which the knowledge will emerge. Further, multi-agent decision-making capability will be helpful in deciding the best action to take. Multi-agent is extended to Reinforcement learning so that policy-based decisions can be made. Figure 21.3 shows the layout of the outlier correction multi-agent architecture. master–slave multi-agent design pattern is used like the previous research papers.

The outlier correction coordinator agent is the master agent in this outlier correction sub-agent. The outlier correction sub-agent receives missing values imputed, and outlier detected time series data. To facilitate regression techniques to predict the outlier correction value, seasonal data for day, hour, week, and month data is calculated. Calculated seasonal data will be sent to the regression sub-agent, which can be used to extend for different types of regression techniques. Currently, it is limited to six algorithms but this can be extended to different regression techniques. Similar to missing value imputation, restartable agents are used. Once the best regression technique is selected, an outlier fixing agent will be initiated. Unlike missing values, outlier values are limited. Therefore, primarily, the restatable agent feature is not required.

Outlier correction is an iterative process, as fixing of a few outliers may result in new outliers. Therefore, outlier correction should be done until the number of outliers is reduced, and the optimum number should be selected.

To implement a decision-making and knowledge-emerging platform, after iteration Table 21.1, data will be collected as similar to missing value imputation and outlier detection sub-agent.

21.4 Multi-agent Implementation

Since human mobility is state space data, this data is converted to time series. State space data was converted to ten locations. Since time series forecasting requires continuity data, missing values were imputed by using the multi-agent architecture. Rule-based and regression techniques were used for the mission values imputation. Out of 83,000 records, around 4809 records (5.79%) were imputed. Once the missing data was imputed, pre-processed data was sent to the outlier detection.

. The best outlier fixing regression technique was identified using the harmonic mean of the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Relative Squared Error (RAE).

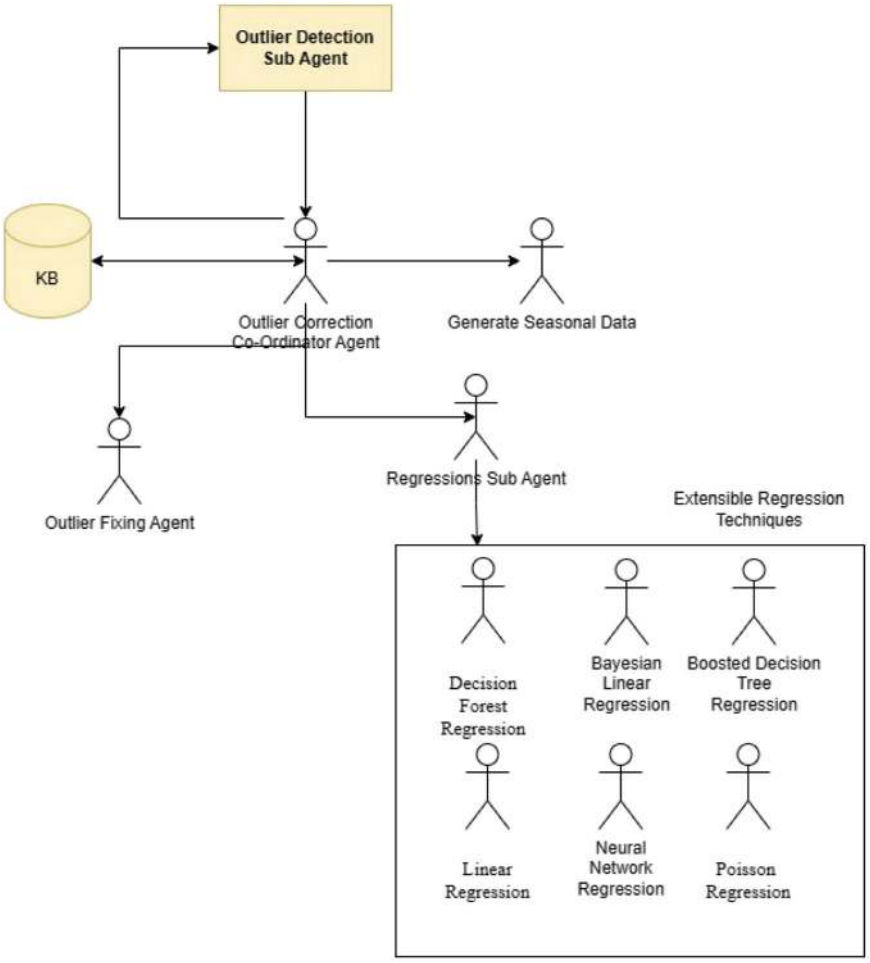


Fig. 21.3 Layout of the outlier correction multi-agent architecture

Table 21.2 shows the evaluation parameters and harmonic mean for the Hiroshima Botanical Garden. To differentiate the best regression technique, harmonic was calculated to the 4th decimal point and the least value with the regression technique is chosen as the best technique for the location. Since data was pre-processed for ten locations in Hiroshima, Japan, the same technique was used to identify the best technique for each location as shown in Table 21.3. It is noticed that the boosted decision tree regression is suited for public places.

For the given location, Hiroshima Botanical Garden twelve iterations were executed by the sub-agent as shown in Fig. 21.4.

For the given location, Hiroshima Botanical Garden, 12 iterations were executed by the sub-agent as shown in Fig. 21.4. As seen in Fig. 21.4, iteration eighth is the

Table 21.1 Outlier fixing techniques per location

Data point	Values
Data type	Mobility/temperature
Sub-data type	Location type: public place, transport-related, office
Number of records	0–1,000/1,000–100,000/100,000>
Frequency	Hourly/Daily/Monthly
Number of outliers	0–100/100–1,000/1,000>
IsTrend	Yes/no
IsSeasonal	Yes/no
Regression technique	Decision forest regression Bayesian linear regression Boosted decision tree regression Linear regression Neural network regression Poisson regression
Number of iterations	0–5, 6–10, 10>

Table 21.2 Regression value comparisons for Hiroshima botanical garden

Technique	MAE	RMSE	RAE	RSE	Harmoic mean
Decision forest regression	0.70	0.94	0.78	0.71	0.7718
Bayesian linear regression	0.70	0.92	0.79	0.68	0.7616
Boosted decision tree regression	0.71	0.92	0.78	0.68	0.7622
Linear regression	0.70	0.92	0.79	0.68	0.7616
Neural network regression	**0.71**	**0.91**	**0.79**	**0.67**	**0.7596**
Poisson regression	0.72	0.95	0.81	0.72	0.7897

Table 21.3 The best outlier fixing techniques per location

Location	Regression technique
Ballpark	Boosted decision tree regression
Bomb site	Linear regression
Hachobori, Hiroshima	Boosted decision tree regression
Hiroshima botanical garden	Neural network regression
Hiroshima city hall	Boosted decision tree regression
Hiroshima MOCA	Boosted decision tree regression
Hiroshima prefectural office	Boosted decision tree regression
Hiroshima station	Boosted decision tree regression
Kamiyacho, Hiroshima	Boosted decision tree regression
Hiroshima castle	Boosted decision tree regression

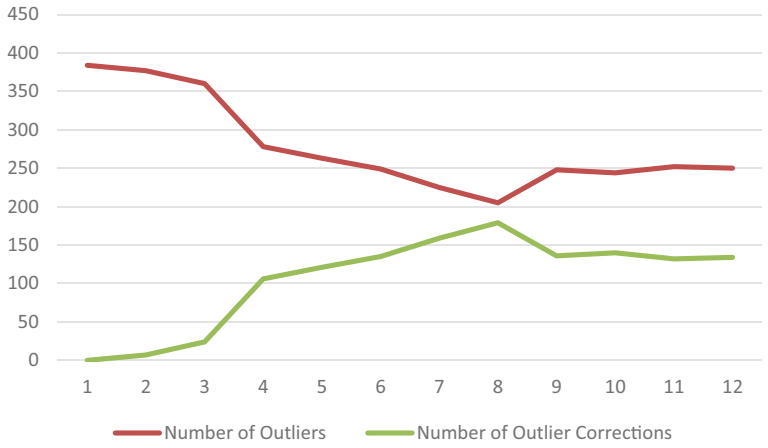


Fig. 21.4 Outlier corrections during each iteration

most optimal iteration where more than 45% outliers were fixed. Table 21.4 shows the outlier fixing for each of the ten locations.

Table 21.4 Outlier fixing for all locations

Location	Total points	Total outlier points	Optimum outlier fixes	Optimum iteration	Percentage (%)
Ballpark	8785	596	276	6	46
Bomb site	8784	898	290	5	32
Hachobori, hiroshima	8785	1130	623	10	55
Hiroshima botanical garden	8736	384	179	8	47
Hiroshima city hall	8785	662	345	7	52
Hiroshima MOCA	8785	296	127	5	43
Hiroshima prefectural office	8785	792	423	8	53
Hiroshima station	8785	1021	578	11	57
Kamiyacho, hiroshima	8785	1073	590	11	55
Hiroshima castle	8785	315	120	4	38

As shown in Table 21.4, the sub-agent has fixed the outliers for each location. Even though complete outlier correction is impossible, outliers were corrected to achieve the best forecasting possible.

One of the key reasons to use multi-agent techniques is to utilize the decision-making and knowledge-emerging features. In this research, policy-based reinforcement learning is used for decision-making and knowledge-emerging features.

21.5 Concluding Remarks

This research has proposed a master–slave multi-agent technique with the coordinating agent to outlier correction to forecast time series. As this research is a part of the AutoML platform for time series forecasting, multi-agent techniques were used for decision-making and knowledge-evolving processes. Once the outliers were detected, a better regression technique was selected using the harmonic mean of regression evaluation parameters. Due to the nature of outliers, multiple iterations were executed until the optimum number of iterations was identified. It was noticed that it is impossible to correct all the outliers, therefore its target is to fix maximum outliers to improve the forecasting of time series.

In the future, the AutoML platform will be introduced with reinforcement learning for decision-making. As the AutoML platform is developed for pre-processing, missing value imputation, outlier detection, and correction, the next is to extend the AutoML platform for normalization and forecasting.

References

1. Abdoola, F., Kruse, C.: Evaluation of Google AutoML Vision artificial intelligence in detecting referable diabetic retinopathy in South Africa. SA Ophthalmol. J. 17–21 (2023)
2. Asanka, D.: AutoML in Azure machine learning for regression and time series. (SQLShack) (2021). https://www.sqlshack.com/automl-in-azure-machine-learning-for-regression-and-time-series/. Accessed 16 Aug. 2021
3. Asanka, P.D., Karunananda, A.S.: Troubleshooting in software as a service (SaaS) environments using multi-agent technology. Int. J. Innov. Res. Technol. **8**, 1 (2014)
4. Asanka, P.D., Rajapaksha, C., Takahashi, M.: Automated missing value imputation in time series using multi-agent technologies. IEEJ (On Review)
5. Asanka, P.D., Rajapakshe, C., Takahashi, M.: Identifying unusual human movements using multi-agent and time-series outlier detection techniques. In: ICARC 2023. University of Sabaragamuwa, Belohuloya, Sri Lanka (2023)
6. *AutoML (Automated Machine Learning).*: (MathWorks) (n.d.). https://www.mathworks.com/discovery/automl.html. Accessed 8 Feb. 2023
7. *AutoML.org.*: (n.d.). https://www.automl.org/automl/auto-sklearn/. Accessed 8 Feb. 2023
8. *EvalML.*: (n.d.). https://evalml.alteryx.com/en/stable/index.html. Accessed 5 Feb. 2023
9. Kazijevs, M., Samad, M.D.: Deep imputation of missing values in time series health data: a review with benchmarking (2023). https://doi.org/10.48550/arXiv.2302.10902

10. Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K.: Auto-WEKA 2.0: automatic model selection and hyperparameter optimization in WEKA. J. Mach. Learn. Res. **17** (2016)
11. Kumar, S.: 4 techniques to handle missing values in time series data (2022). https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-series-data-c3568589b5a8. Accessed 19 Mar. 2023
12. Li, Q., Li, R., Ji, K., Dai, W.:. Kalman filter and its application. In: 2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS). Tianjin, China (2015). https://doi.org/10.1109/ICINIS.2015.35
13. *Machine Learning in your Database using SQL*.: (MindsDB) (n.d.). https://mindsdb.com/. Accessed 28 Aug. 2021
14. Pratama, I., Permanasari, A.E., Ardiyanto, I., Indrayani, R.: A review of missing values handling methods on time-series data. In: 2016 International Conference on Information Technology Systems and Innovation (ICITSI). Bandung-Bali, Indonesia (2016)
15. Sokannit, P., Chujai, P.: Forecasting household electricity consumption using time series models. Int. J. Mach. Learn. Comput. **11**(6) (2021)
16. *Tpot AutoML*.: (n.d.). https://www.geeksforgeeks.org/tpot-automl/. Accessed 8 Feb. 2023
17. Yin, H., Liu, K.-H., Sun, M., Chen, Y.: AutoML for large capacity modeling of meta's ranking systems. In: ACM Web Conference 2024, New York, USA (n.d.). https://doi.org/10.1145/nnnnnnn.nnnnnnn

Chapter 22
Similitude Assessment Using Iramuteq® Considering the Triad, Corporate Governance, and the Risk

Samira Sestari do Nascimento and Jair Minoro Abe

Abstract Corporate governance is a set of processes and regulations for more effective managerial control in organizations. The system is distinct, considering the phenomena that can interfere in the organizational dynamics. In textual analysis of corporate governance, risk, and the triad, the IRaMuTeQ software was used to analyze the correlation. The objective is to analyze the similitude by mapping the correlation between the articles. The research is qualitative, combining with a specific methodology of dynamic similitude. A textual corpus was used considering the history and chronology of the facts. The participants are four researchers with experience in analyzing inconsistencies of propositions and facts, which contributed to the discussions and final considerations. The articles were searched in the main scientific databases using the filter method, chronology of facts. In the software, the baseline analysis was classification and multivalued analysis of similitude through tree correlation. The results of the factual analysis of the graphs were visually demonstrated in the tree the correlation between decision and risk; structure and control; and the adjacent ones such as the corporate and company, presented independence, that is, there is no correlation. The result of the study demonstrates a risk-oriented view of decisions and governance. The procedure of analysis of the corpus text evidenced the correlation accordingly and we concluded that the IRaMuTeQ® has a connection between the facts and a logical structure in qualitative studies.

22.1 Introduction

A country's economic development depends on its level of business activity, as companies seek operational efficiency to maximize profits, attract investors, and achieve market projections. The study explores the chronological relationship between the theory of the firm, consultancy relationships, corporate performance evaluation, decision-making groups, and corporate governance theory.

S. S. do Nascimento · J. M. Abe (✉)
Paulista University, R. Dr. Bacelar, 1212, São Paulo 04026-002, Brazil
e-mail: jair.abe@docente.unip.br

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_22

22.2 Theoretical Reference

The study of the reference covers three literary currents that revealed themselves as an ideological triad, being composed of a vision that involves thinkers, reports, the Sarbanes & Oxley law, Corporate Governance attributing to risk; being composed of a vision based on a historical and chronological basis, honoring current authors and studies. The themes were considered immutable.

The OECD Committee Evaluates Public and Economic Policies supporting representative democracy and the rules of the market economy. It focuses on improving international work, promoting a new economic dimension, encouraging more liberal trade policies, and promoting economic development, sustainability, and financial stability in member countries. The OECD is an intergovernmental economic organization comprising 38 strategic member countries, aiming to develop practices to stimulate economic development and social well-being [1].

The Commission's mission focuses on facilitating and resolving common trade problems in the Eastern world through strategic cooperation, analyzing the US economic assistance program for After War II Eastern Europe, aimed at economic reconstruction and the essential role of government in the growth of the private sector; with the objective based on economic reconstruction, the idea of the "Marshall Plan" [2]. In this way, the OECD declares that the government is crucial for the economic growth of the private sector, which depends on increased investments, efficiency in the capital market, and company performance.

The Federal Securities and Exchange Commission (SEC) [3], is an independent agency responsible for protecting and regulating capital markets in the USA, focusing on corporate governance, decision-making processes, and record security. The term "corporate governance" first appeared in the Federal Register, addressing management accounting issues as part of its regulatory purview. The SEC committee, responsible for ethical actions, discusses the relationship between the corporate world and society [4].

Researchers Jensen and Meckling introduced a new domain of corporate governance research, approaching the concepts of the firm theory concerning the issue of control and investigating the nature of the costs generated by the existence of debt with external equity. Other authors explore the concepts of agency costs concerning control, investigating the nature of the costs generated by the existence of debt with external resources [5]. However, Macnulty [6], later demonstrated that corporate governance (CG) analysis directly combines qualitative and quantitative predictions, controlling processes, and related to the organization's planning and decision-making, allowing the evaluation of objectives and performance. Vance [7], examines the structure and performance of top management in corporate governance, revealing that technical experience and internal managerial expertise are key factors for control. The results of the author's study show that about a third of the board is made up of technical knowledge, while managerial experience represents less than a third.

Governance is closely linked to the management of an organization and its governance model and with stakeholders and other interested parties. After growing in the 1980s, the American economy moved to a corporate governance model influenced by external control, resulting in conflicts of interest and information asymmetry.

Considering the differences in the direction of organizations, the Anglo-Saxon governance model believes that wealth is distributed and securities of value developed, reducing risk to shareholders but requiring less direct market monitoring due to market fluctuations [8]. The importance of the government coordination system has been significantly emphasized through assessments of efficiency in organizational control, with top management committed to best practices in assessing liquidity, integrity, and management of organizational accounts [9].

Corporate governance concerns led to the publication of three committee reports: Cadbury, Greenbury, and Hampel. The Bank of England selected Lord Cadbury to develop guidelines on corporate governance for executives, directors, and supervisors. The code, known as the Best Practice Code and later known as the Cadbury Report, was published in December 1992. Cadbury's Financial Advisory Committee, also known as the Corporate Governance Committee [10], assessed investors' lack of confidence in the honesty and responsibility of companies, such as the division of responsibilities of the organization's chairman and the appointment of an audit committee [11], Assigning responsibilities, agents in the company's life are people interested in a specific decision, including influencers and those affected by it executives make organizational decisions maximizing shareholder equity [12]. Cadbury's proposals involve extensive monitoring, evaluation, and control to improve the information provided to shareholders and organizational agents, ensure company control, and strengthen shareholder interests. Forker [13] states that development methods and internal controls must be actively and confidently adopted when legislation, regulations, court decisions, or initiated legal changes are imposed.

Other precursors on the committee, Hampel produced a document providing principles and guidelines for implementing the combined Cadbury, Greenbury, and Hampel recommendations of June 1998 [13]. New versions establish good corporate governance practices, such as the remuneration of directors and remuneration committees, according to the Greenbury Report [14]. Other reports were developed to review the recommended practices and conduct and improve them, with emphasis on the Hampel Report [14], the Turnbull Report [15], and the Higgs Report in 2003. Specific reports are needed for independent auditing, recruitment, and training of external directors to consolidate good governance practices [16]. Specific aspects for financial institutions Walker Review of Corporate Governance of UK Banking Industry [17]. Furthermore, in July 2010, the Financial Aspects of Corporate Governance (FRC) [18], published the UK Stewardship Code, a corporate governance code for institutional investors [19].

On July 30, 2002, Senator Paul Sarbanes and Representative Michael Oxley introduced legislation that requires corporate governance in public companies to focus on financial management to prevent fraud and maintain market balance [20].

The law was enacted to prevent fraud and financial insecurity for investors, aiming to deter fraud and ensure better governance over the information disclosed [21].

When drafting the Sarbox Law or SOx [22], it considered audit and security mechanisms in companies, including rules and responsibilities to supervise activities and operations while reducing risk and transparency. For assurances, the Public Company Accounting Oversight Board (PCAOB) [23], was created to oversee the auditing activities of the (SEC), ensuring compliance with obligations and establishing quality control standards and auditing standards. Directors, especially financial ones, gain an important role in SOx, as the Sarbanes–Oxley Act makes them responsible for monitoring internal controls concerning the disclosure of financial information. It is based on three pillars, including the board of directors, Audit Committee, and Administration Executive.

The board of directors is responsible for hiring or appointing the organization's president, ensuring its stability, controlling structure, and accounting compliance. They also supervise the company's continuity, adhere to ethical standards, disclose necessary financial information to the market, and verify the veracity of this information [24]. The audit committee recommends hiring a financial specialist to evaluate and monitor the internal controls and actions the auditors evaluate. The administration is made up of the executive director, president, and other directors of the company. Internal control and accounting must be implemented appropriately. The annual report must be used to disclose all outstanding financial information. SOx requires an intelligent rethinking of guidelines, new technologies, and cost–benefit ratios [25].

Corporate Governance (CG) centralizes the regulation of corporate structures, organizing powers, and establishing rights and duties of shareholders. The Anglo-Saxon model is based on decentralizing shareholder control and separating property and management rights [26]. Andrade and Rossetti [27], the decentralization of share control and the separation of property and management rights are the primary basis of the Anglo-Saxon model. The Japanese-German model is focused on actions driving the company's strategy, with indicators of performance, activity, and distribution of dividends, demonstrating the efficiency of corporate social policies and their sustainability.

Corporate governance in the US has the most liquid capital market in the world, with the most significant volume, market capitalization and listed companies. The Cadbury Report, published in 1992, is a significant milestone in corporate governance, influencing company management processes, particularly in the United Kingdom, and contrasting with the Anglo-Saxon model.

Companies' capital is concentrated, with financing coming mainly from banking sources. Corporate governance models in Latin America, based on family and state-owned companies, are similar in Italy, France, Spain, and Portugal, with less clear sources of financing [28]. The Brazilian governance model focuses on the management of an organization and its relationship with shareholders and other stakeholders, such as customers, employees, suppliers, and the community. Law for Brazil no. 6,404 [29], establishes that companies or companies with limited liability will divide their capital into shares, with shareholders' liability limited to the price of the resulting shares. The law describes the controlling shareholder's corporate governance duties and rules, emphasizing his responsibility to the company's other shareholders, the community in which he works, and the company's rights and interests."

The objectives and requirements of the law require companies to be transparent about their structures, risk management practices, internal control, and administration composition [30].

The basis of GC is made up of four essential elements: transparency, explanation of actions, consequences, and diligence in attribution, as well as the ability to provide relevant information to interested parties. The growth of the capital market is driven by institutional investments, financing business projects, and necessitating a transition to sustainable capitalism, reflecting the global cultural revolution in search of transparency and ethics [31].

Corporate responsibility involves governance agents in ensuring the organization's economic-financial forecast, reducing financing restrictions for operations, increasing positivity, and considering finances, capital, and human resources [32].

The Brazilian Institute of Corporate Governance—IBGC [33], identifies criteria for companies to adopt good corporate governance practices, including assessment, access to low-cost markets, consolidation of international financial markets, support for strategic alliances, alignment between stakeholders, alignment of interests, reduction of conflicts, improvement of management processes and improvement of corporate image. Recommends actions that directly demonstrate the company's financial situation, risk management, and sustainability monitoring.

The Coso Committee of Sponsoring Organizations of the Treadway Commission [34], suggests that decision-making can be based on risk management frameworks, internal control assessment, fraud detection, and deviation reduction. Organizations focus on good practices in internal control, leadership, and organizational strategy, involving coordination between departments and public bodies. Understanding risk is an art of decision-making that affects credibility, relevance, and trust in modern economies, demanding greater transparency, responsibility, and leadership in risk management [35].

The risk management policy is based on best practices, guiding organizations in understanding internal controls and ensuring that financial reports are reliable to prevent fraud. In 2004, COSO published the Enterprise Risk Management—Integrated Framework (ERM) or COSO II, focusing on enterprise risk management. The ERM—Integration with Strategy and Performance version adds risk principles to strategy and performance improvement [36, 37]. Enterprise Risk Management Integrated with Strategy and Performance [35] changed to protect and enhance stakeholder value, with the governance philosophy in conjunction with corporate governance and with responsibilities and obligations related to the transparency of financial information. Risk management is crucial for companies, ensuring internal control and the reliability of financial reports. Transparent corporate governance reduces uncertainty for investors. Objectives must be aligned with the organizational mission, vision, and values, promoting performance in decision-making.

22.3 Methodology

The research uses a qualitative approach to discover structured and chronological contexts, focusing on the non-quantifiable reality within meanings, motivations, aspirations, beliefs, values, attitudes, and knowledge. The study procedure is divided into two parts, with the first stage analyzing the data in sub-steps, identifying central themes such as the triad, corporate governance, Sox Act and corporate risk relationships in Fig. 22.1.

The search for keywords in scientific databases follows the second stage. The filter is established by title, contains; "keyword", refined by peer-reviewed journals; cite topics in the bases: JSTR; Sage Journal; Elsevier; Science Web; and Wiley Online Liberty. The words were separated by the connective "AND" being "theory of

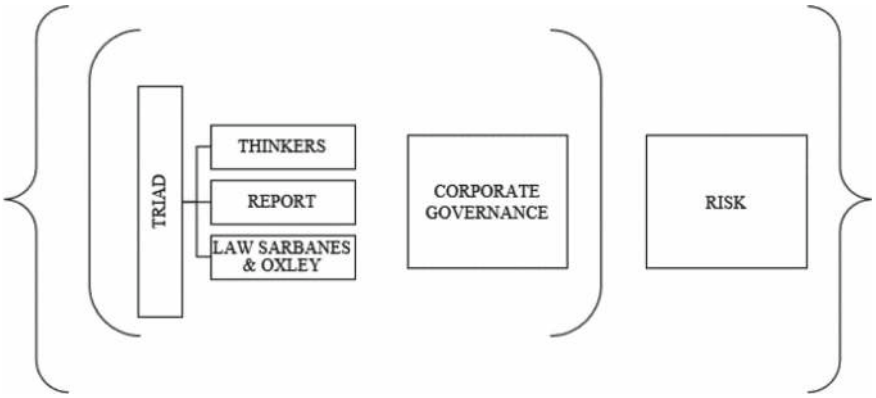


Fig. 22.1 The relationships. *Source* Authors

the firm", "Organizations", "Jensen A"; "Meckling", "Report"; "Carbura Committee", "Sarbanes Oxley Law", "Corporate Governance"; "COSO" and "Organizational risks". Considering the central theme, chronologically considered (1960–2011). The system used to process the metadata of the similarity analysis and summaries assisted by the Interface de R program pour les Analyzes Multidimensionally de Texts et de Questionnaires (Iramuteq®), to extract the "similarity", its connection between words helping by textual corpus. Similarity is a representation based on graph theory and allows you to identify occurrences and connections between words; its results assist in the structured identification of a textual corpus and its relationships. The instrument is a free and open-source program created by Pierre Ratinaud for statistical analysis of textual corpus and tables based on their word composition. After reading the article summaries, the legitimacy and chronology of the facts were considered. The selection filters were eligibility; context; inclusion and exclusion of articles. The second part of the study, which includes the textual analysis of article summaries, considers in Table 22.1.

Table 22.1 Textual analysis of the articles

Nº	Articles	Description	Chronology
1	Theory of the firm: managerial behavior, agency costs, and ownership structure	The thinkers	1976
2	Who controls whom? An examination of the relationship between management and boards of directors in large American corporations		1977
3	Corporate governance: assessing corporate performance by boardroom attributes		1978
4	Group decision-making and communication technology		1992
5	Corporate governance and disclosure quality		
6	Corporate governance: from accountability to enterprise		1999
7	Enron: what happened, and what can we learn from it? Journal of Accounting and Public Policy	Law Sarbanes–Oxley	2002
8	Does "Good" Corporate Governance Help in a Crisis? the impact of country- and firm-level Governance Mechanisms in the European Financial Crisis		2012
9	Corporate governance and corporate performance: a comparison of Germany, Japan, and the U.S.		1997
10	An investigation of the relationship between self-serving attributions and corporate governance		2009
11	The History of Corporate Governance		2011
12	A conceptual framework for corporate risk disclosure emerging from the Agenda for corporate governance reform	Risk	2000
13	COSO ERM: integrating with strategy and performance		2018

Source Authors

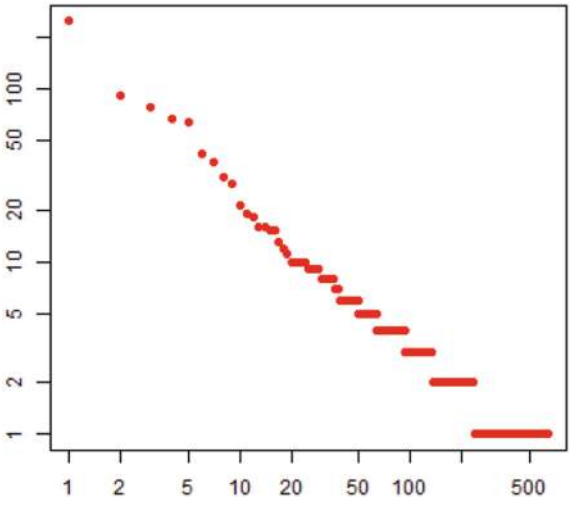


Fig. 22.2 Result textual statistic. *Source* Marchand and Ratinaud [39]

The third part submits a transcription of "Copus_test_03," presenting summaries of analyzed bibliographies in a Word Frequency Diagram, displaying logarithms of weights on the abscissa and ordinate axes. The meaning of y (0; >100) and on the x-axis (0; >500), in summary, follows the meaning of each occurrence: the number of texts: is the number of texts (records) that are contained in the corpus; the number of occurrences: is the total number of words contained in the corpus; the number of forms present in the corpus (active words); the number of hapaxes: number of words that appear only once in the entire corpus and the average of occurrences per text: (number of occurrences)/(number of texts), according to Fig. 22.2.

In the graph, we have the logarithms of the "weights" on the abscissa axis (position of word frequencies in descending order) and the form frequencies on the ordinate axis. Following the analysis of specificities and C.F.A., it associates texts, that is, enabling the analysis of textual production in terms of characterization variables, considering the "active" forms used and selecting by "modalities" of the articles between "ARTICLE_1" TO "ARTICLE_13". The database is divided according to the quadrants and dimensions between the axis (0, 0); below lower correlation between words and above with more excellent correlation, according to the characteristic of a species (Corpus) according to Fig. 22.3.

In the correlation tree, according to Fig. 22.4, produced with the connection with the articles according to Image 3, it generated four nuclei of words. In the correlation tree, the main core (central) has the word: "corporate governance", interconnecting with the other cores with the core words: company, control, structure, risk, and decision; what between the association pairs, a stronger relationship between the words is identified by the quadrant: "Risk and decision" as represented in the correlation tree Fig. 22.4.

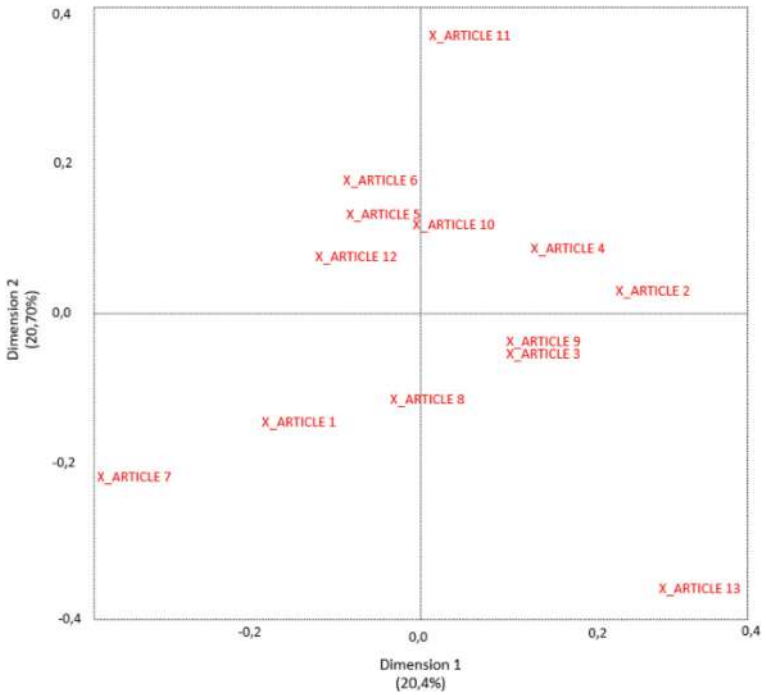


Fig. 22.3 Result textual statistic of the article. *Source* Marchand and Ratinaud [39]

Based on Fig. 22.4, considering the graphic particularities, the central nexus of corporate governance remains with the close connection of facts between control, management, and risk. What defines the facts are in similarity, grouped together. The company group indicates a strong influence between the facts between performance and director, emphasizing the need for continuous corporate governance actions. The decision group is disconnected due to the structure, which interferes with understanding the facts chronologically after understanding governance in organizations.

Therefore, broader studies are necessary to establish generalizations about the topic investigated. The Iramuteq system, despite its limitations, provided similarity analysis considering descriptive, chronological analysis, and their influences.

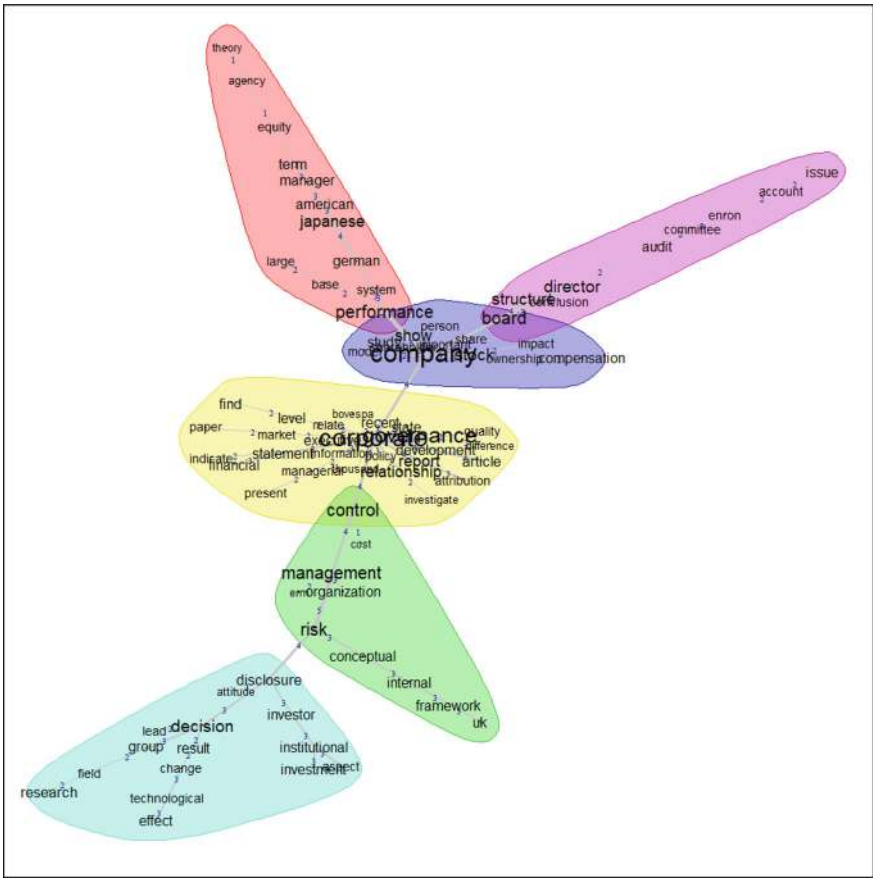


Fig. 22.4 Result similarity map. *Source* Marchand and Ratinaud [39]

References

1. du CEPES, G.A.: Bilan et perspectives de l'intégration économique européenne Politique etrangere. Politique etrangere, pp. 48–56 (1961)
2. OECD.: The "Marshall Plan" speech at Harvard University (1947)
3. Securities and Exchange Commission.: Abrief summary of financial proposals filed with and actions by the S.E.C (1970)
4. Aldrich, H.E., Pfeffer, J.: Environments of organizations. Ann. Rev. Sociol. **2**, 79–105 (1976)
5. Jensen, M.C., Meckling, W.H.: Theory of the firm: managerial behavior, agency costs and ownership structure. J. Financ. Econ.**3**, 305–360 (1976)
6. MacNulty, C.A.R.: Scenario development for corporate planning. Futures **9**, 128–138 (1977)
7. Vance, S.C.: Corporate governance: assessing corporate performance by boardroom attributes. J. Bus. Res. **6**, 203–220 (1978)
8. Coffee, J.C.: Shareholders versus managers: the strain in the corporate web. Mich. Law Rev. **85**, 1–109 (1986)
9. Monks, R.A.G., Minow, N.: Power and accountability. Robert Monks at Stephanie (1991)

10. Cadbury, A.: Report of the Committee on the Financial Aspects of Corporate Governance. Gee, London (1992)
11. Baker, M.B.: Private codes of corporate conduct: should the fox guard the henhouse. U. Miami Int.-Am. L. Rev. **24**, 399 (1992)
12. Kiesler, S., Sproull, L.: Group decision making and communication technology. Organ. Behav. Human Decis. Process.**52**, 96–123 (1992)
13. Forker, J.J.: Corporate governance and disclosure quality. Account. Bus. Res. **22**, 111–124 (1992)
14. Tunc, A.: Le gouvernement des sociétés anonymes au Royaume-Uni: le rapport du Comité Hampel. Revue internationale de droit comparé **50**, 912–923 (1998)
15. Short, H., Keasey, K., Wright, M., Hull, A.: Corporate governance: from accountability to enterprise. Account. Bus. Res.**29**, 337–352 (1999)
16. Turnbull, N.: Internal Control: Guidance for Directors on the Combined Code: known as the Turnbull Report. CIMA, London (1999)
17. Murcia, F.D.-R., De Souza, F.C., Borba, J.A.: Continuous auditing: a literature review. Revista Organizações em contexto **4**, 1–17 (2008)
18. Walker, D., others.: A review of corporate governance in UK banks and other financial industry entities (2009)
19. FRC Financial Reporting Council.: The UK Corporate Governance Code (2010)
20. Tarraf, H.: Literature review on corporate governance and the recent financial crisis (2010). SSRN 1731044
21. van Essen, M., Engelen, P.-J., Carney, M.: Does "Good" corporate governance help in a crisis? The impact of country- and firm-level governance mechanisms in the European financial crisis. Corp. Gov.: Int. Rev.**21**, 201–224 (2012)
22. Benston, G.J., Hartgraves, A.L.: Enron: what happened and what we can learn from it. J. Account. Public Policy **21**, 105–127 (2002)
23. Sarbanes, P.S.: Lei Sarbanes-Oxley-Lei Sox. American Institute of CPA-AICP*A*. Recuperado de (2002). https://www.aicpa.org
24. Public Company Accounting Oversight Board.: Annual report (2022)
25. Mizruchi, M.S.: Who controls whom? An examination of the relation between management and boards of directors in large American corporations. Acad. Manag. Rev. **8**, 426–435 (1983)
26. Deloitte Touche Tohmatsu Limited.: (Re-think) SOx Compliance (2022)
27. Prowse, S.: Corporate governance: Comparaison internationale. Une étude des mécanismes de contrôle d\textquotesingleentreprise aux États-Unis, en Grande-Bretagne, au Japon et en Allemagne. Revue d\textquotesingleéconomie financière**31**, 119–158 (1994)
28. Andrade, A., Rossetti, J.P.: Governança corporativa: fundamentos, desenvolvimento e tendências. Atlas São Paulo (2004)
29. Kaplan, S.N.: Corporate governance and corporate performance: a comparison of Germany, Japan, and the US. J. Appl. Corp. Financ. **9**, 86–93 (1997)
30. Lei nº 6.404.: Dispões sobre as Sociedades e Ações (1976)
31. da Luz, A.T.M., Pagliarussi, M.S., Teixeira, A.M.C., Baptista, E.C.: An investigation of the relationship between self-serving attributions and corporate governance. Braz. Bus. Rev.**6**, 181–197 (2009)
32. Rudge, L.F.: Mercado de capitais. Elsevier Brasil (1993)
33. Cannon, T.: Corporate Responsibility: a Textbook on Business Ethics, Governance, Environment: roles and Responsibilities (1994)
34. Instituto Brasileiro de Governança Corporativa.: Código das Melhores Préticas de Governança Corporativa 5ª (2015)
35. COSO Committee of Sponsoring Organizations of the Treadway Commission.: Enterprise risk management (2017)
36. Solomon, J.F., Solomon, A., Norton, S.D., Joseph, N.L.: A conceptual framework for corporate risk disclosure emerging from the agenda for corporate governance reform. Br. Account. Rev.**32**, 447–478 (2000)

37. Djasuli, M., Triyuwono, I., Purwanti, L., Roekhudin, R.: Committee of sponsoring organization of the treadway commission (COSO) framework as a control framework construction internal sharia based. Budapest Int. Res. Critics Inst.-J. (BIRCI-J.) **5** (2022)
38. Sobel, P.J.: COSO ERM: integrating with strategy and performance. In: The Institute of Internal Auditors International Conferrence, UAE, Dubai, vol. 6 (2018)
39. Marchand, P., Ratinaud, P.: L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT, vol. 2012, pp. 687–699 (2012)
40. OECD.: List of OECD member countries—Ratification of the convention on the OECD (1960)
41. Bortolon, P., Junior, A.: Delisting Brazilian Public companies: empirical evidence about corporate governance issues. Braz. Bus. Rev.**bbrconf**, 92–117 (2015)
42. Cheffins, B.R.: The history of corporate governance. SSRN Electron. J. (2011)

Chapter 23
Logistical Challenges in Last-Mile Deliveries in the Outskirts of the City of São Paulo. An Analysis of the Cargo Theft and Robbery Rates of an E-commerce Company

Kennya Vieira Queiroz and Jair Minoro Abe

Abstract The logistics of transportation and distribution of goods is considered one of the most critical stages in the production chain of different product, and service segments worldwide. Urban freight distribution is a vital function for the flow of production, the delivery of goods, and the provision of services to society. However, large urban centers face many logistical problems that affect the delivery and distribution of goods. Some examples include the lack of urban road infrastructure, high operating costs, and lack of planning for roads and highways, among other factors that affect the entire logistics chain. Another severe problem related to the distribution of goods is the high rates of cargo theft, especially in significant suburbs. In this scenario, it is crucial to evaluate the financial losses resulting from thefts in an e-commerce company in the city of São Paulo—Brazil, seeking to understand and quantify the financial and material losses associated with cargo theft. The research adopts a quantitative–qualitative or mixed approach. Its theoretical support comes from the following authors: Araújo (2012), Ballou [1], Manerba [12], documents from the websites of the National Transportation Confederation-CNT (2009), Data Favela, and the Central Única das Favelas-CUFA [2, 3], among others; based on studies carried out in 2023. The empirical results of the analysis reveal that in the year 2023, it was found that the eastern region of Greater São Paulo had the most significant loss from cargo theft; out of the company's total losses of R$279.536,48, the eastern region had R$122.266,09. This amount is equivalent to almost 44% of all losses, with a high concentration in April, September, October, and November. Hence, there is a need for the company to have good data management and analysis, as well as risk management, and to develop actions to combat cargo theft to identify the relevant variables employing strategies to reduce thefts.

K. V. Queiroz · J. M. Abe (✉)
Paulista University, Rua Doutor Bacelar, 1212, Bairro Vila da Saúde, São Paulo, SP, Brazil
e-mail: jairabe@uol.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_23

23.1 Introduction

E-commerce refers to online business, including buying and selling products and services. These transactions are carried out using electronic devices such as computers, cell phones, or tablets.

Laudon and Laudon [8, p. 117] states, "Companies can use the Web to offer uninterrupted information, services, and customer care, creating positive interactions with customers that serve as the basis for longer-lasting relationships and regular purchases".

According to the author, this type of service originated around 1960 in the United States, with the sole aim of exchanging order request files to inform companies whether customers were interested in ordering the product.

The E-Commerce service, as it is known today, was given a boost in the 1990s when Amazon and eBay began to show interest in the system, but it was not until 1999 that virtual commerce took off with the purchase of the Booknet store, founded by writer and economist Jack London, which was renamed Submarino. From then on, this movement was accepted by consumers and became popular, especially in more developed countries.

Some others still had the traditional way of buying and selling, where the retailer was responsible for buying the goods and selling them in the store, the consumer was responsible for taking the product home when they bought it, and distribution was only the responsibility of the retailers when it came to products with a larger volume. According to Fleury and Lana (2000), "Only chain stores worked with a distribution system, but it was the one between the distribution center and each of the stores—in which trucks left at a certain frequency, time and load, heading for half a dozen fixed addresses—a completely different process from individualized delivery".

Today, we have many E-commerce, including Amazon, Shopee, Mercado Livre, and Madeira Madeira. According to Novaes et al. [16]:

> In the case of e-commerce, essentially B2C (business-to-consumer) marketing, the potential customer, driven to buy via the Internet for various reasons, attaches great importance to logistical factors, causing the company to pay special attention to the logistical infrastructure, combining stocks, distribution, information processing, and human resources.

E-commerce has grown a lot in Brazil in recent years, and the lockdown period caused by the COVID-19 pandemic was one of the factors that significantly influenced this growth and the change in Brazilians' shopping culture.

E-commerce is short for electronic commerce. It grew more robust with the arrival of the Internet, making the whole buying and selling process more manageable. At first, only small products were sold, such as CDs, DVDs, books, etc. Nowadays, cars, houses, yachts, aeroplanes, works of art, and other luxury products are sold [5, p. 2].

Even with the end of the pandemic, consumer behavior has remained, causing the culture of using e-commerce to make purchases, causing e-commerce companies to need to develop and optimize their processes in order to promote a safer purchase and delivery service, meeting customer expectations and maintaining the growth of this sector.

> Customer services are the essential elements within the concept of order cycle time that the logistics professional can control. Order cycle time is defined as when the customer places the order and when the product is delivered. The order cycle is the sum of all the time events that can be measured in the total time it takes to deliver an order [1].

Therefore, this is a complex cycle that deals with receiving, stocking, controlling orders, the delivery process, tracking orders sent, the performance of carriers that are responsible for deliveries with structuring and organizing these processes, and logistics, which can be a company's most significant positive differential in e-commerce, promoting its credibility and evolution.

23.2 Theoretical Framework

23.2.1 The Outskirts

Favelas, often neglected by economic groups and companies in Brazil and sometimes considered risk areas for e-commerce delivery operations, represent important centers of life and culture, home to an estimated 17.1 million people living on the outskirts of Brazilian cities, according to research by the Locomotiva Institute in partnership with Data Favela and the Unified Slum Center (CUFA).

Despite this numerically significant population, operations involving the delivery of products in these areas face significant difficulties.

Araújo (2012) states, "Urban distribution has become one of the great challenges facing businesses". Brazil's major cities have been experiencing problems related to a lack of planning and road infrastructure, traffic jams, a lack of suitable places for unloading, disorderly growth, a lack of organization of the regular addressing of properties, an inadequate architectural structure for receiving the product, high rates of theft and crime, among others. All of these logistical challenges have required a specialized approach and structured planning.

According to Data Favela [2], 13.6 million people are living in favelas in Brazil, and they generate an annual turnover of R$119.8 billion; according to the same study, 39% of residents shop online, with younger people making up the most significant proportion with regular internet access (over 97%), and 87% of adults accessing the Internet at least once a week.

The Brazilian market has grown significantly, according to da Silva et al. [19]. "Digital retail earned R$35.2 billion in the first quarter of 2021, an increase of 72.2% over the same period last year".

Although this is an attractive audience, there is a big challenge for transport companies to reach these places where they often do not have a ZIP code (Postal Address Code) or a residence number and lack urban infrastructure and security, which makes it difficult for goods to reach their destination, thus excluding slum dwellers from delivery services.

23.2.2 Last-Mile Logistics

The concept of the last mile is recently introduced in the logistics literature. Lim et al. [9] state that last-mile logistics is the last leg of a B2C (business-to-consumer) parcel delivery service. It takes place from the point where the orders enter the manufacturing center, distribution center, and store, among others) to the point where the final recipient arrives [6].

According to the authors, although this stage is critical and decisive, it is one of the most critical transportation activities and is considered to be the most expensive, least efficient and most polluting in the entire supply chain [6].

The complexity of last-mile delivery in Brazil is further exacerbated by the lack of adequate infrastructure, such as addresses on unnamed streets or in locations unnumbered; areas without lighting, paving, or sanitation; lack of urban road infrastructure; high operating costs; lack of planning of roads and highways; among other factors; as well as natural risks, such as floods. Another severe problem faced concerning the distribution of goods is related to security risks, such as theft and robbery of cargo, especially in large suburbs, which has affected logistics service providers and carriers [4, 20].

In addition to the factors already mentioned that can affect this service, we also have the moment of evaluation of customer satisfaction, which occurs when the consumer evaluates the delivery time and the conditions in which the package arrived. In this way, the last mile is fundamental for e-commerce, as it is the stage at which the shopkeeper shows the quality of the service.

23.2.3 Factors that Influence the Parcel Delivery Process

Manerba [12] points out that last-mile delivery is currently considered one of the process's most expensive, least efficient, and polluting parts.

With the chaotic infrastructure in large urban centers, coupled with the disorderly growth of large communities without formal addresses and with a frequent increase in risk areas, part of society is not served with the delivery of their e-commerce orders, having no option to pick them up in places that allow access to this shopping model.

The situation has been considered one of the major bottlenecks in e-commerce logistics, with this process causing a considerable amount of damage and losses.

According to Manuj and Mentzer [13], security risk can be characterized as: "a threat from a third party who may or may not be a member of the supply chain and whose motivation is to steal data or goods and/or destroy, disrupt or disable a company's operations".

Also, according to the authors, among the most significant security risks are so-called freight breaches, which are violations of the integrity of cargo and products, leading to the loss or tampering of goods.

Another factor is related to the receipt of the package; it can happen that consumers are not at home when a package is placed on the delivery route to customers, and then a "failure" in delivery occurs [10].

23.2.4 Cargo Theft in Brazil

Cargo theft is another serious problem faced by the distribution of goods, and this is a general problem to be faced by logistics operations, especially concerning road transportation. According to data from the National Transport Confederation—CNT (2009), 61% of all Brazilian cargo is transported by road, contributing to the high rates of cargo theft in Brazil.

For logistics professionals, cargo theft is an even more significant challenge in emerging economies [15]. The most direct consequence of theft is not only the loss of cargo and its financial impact, but the damage caused goes beyond this, such as operational expenses with rework, new freight costs, delays, and loss of service level, thus generating negative impacts on the company's reputation [17, 21].

23.3 Methodology

This study analyzes financial and material losses associated with cargo theft in an e-commerce logistics company.

The methodological process involved in this research is a quantitative–qualitative or mixed nature case study using a combination of qualitative and quantitative methods. Data related to cargo theft was collected from a company that operates in furniture e-commerce in 2023. The process involved exploratory and descriptive research and quantitative analysis of data collected via field research, which aims to identify the losses caused by cargo theft and robbery in an e-commerce logistics company in the city of São Paulo—Brazil.

As presented by Gil [7, p. 43], exploratory research is a model that aims to provide the reader with a proximity to the object of research and its problem, make it explicit, and build hypotheses. The case of this research will involve a bibliographical review to define the concepts that may be in doubt and an analysis of studies of existing practices that generate examples and facilitate understanding, which was carried out by surveying theoretical references published in written and electronic media, such as books, scientific articles and websites.

Descriptive research was also used when the aim was to describe the reality of transportation in risk areas through data analysis and the basis of documentary analysis as recommended (Triviños 1987).

According to Marconi and Lakatos [14], the following stage deals with documentary research, which naturally draws on more diverse and broader sources

without undergoing a prior analytical filter or scientific scrutiny: statistical tables, organizational reports, etc.

As for the nature and technique of data analysis, the research will have quantitative analysis, according to Malhotra [11, p. 154], "seeks to quantify the data and usually applies some form of statistical analysis" and aims for objective clarity, social understanding of the impact of the phenomenon on a population.

Qualitative analysis, on the other hand, is characterized, according to Gerhardt and Silveira (2009, p. 31), by not being concerned "with numerical representativeness, but rather with deepening the understanding of a social group, an organization". This scenario will give us a broader and more contextualized view of the problem and its impact on e-commerce logistics companies.

23.4 Analysis of the Data

Table 23.1 shows the period of occurrence and the number of notes/orders stolen.

As can be seen, of the 360 total occurrences of stolen notes/orders, January had 56, February 30, April 60, May 01, June 08, July 06, August 49, September 47, October 60, November 43. Therefore, April and October recorded the highest occurrence rate of the year. Of particular note was the East Zone, the only region to record robberies in November, when Black Friday sales took place.

The data collected makes it possible to pay greater attention to specific months, looking for strategies to offer more significant support to prevent theft (Table 23.2).

The data shows that in 2023, of the total R$279.536,48 in losses due to theft and robbery at the e-commerce logistics company surveyed, the eastern region of the greater São Paulo Capital was where the most significant losses were concentrated, followed by the southern region (Graph 23.1).

The graph shows that in addition to the eastern region is the region with the highest number of losses in November, a month of extreme criticality for the company because it is the month of Black Friday sales actions when there is an accelerated increase in deliveries; it was the only region in the city of São Paulo that had cargo theft and robbery in that month.

23.5 Results

The data for this research was obtained from an e-commerce logistics company's theft database, specifically at the Last Mile stage of deliveries. The empirical results of the analysis reveal that in the year 2023, it was found that the eastern region of Greater São Paulo had the most significant loss from cargo theft; out of the company's total losses of R$279.536,48, the eastern region had R$122.266,09, equivalent to almost 44% of all losses, with a high concentration in April, September, October, and November. These results show the need for the company to have good data

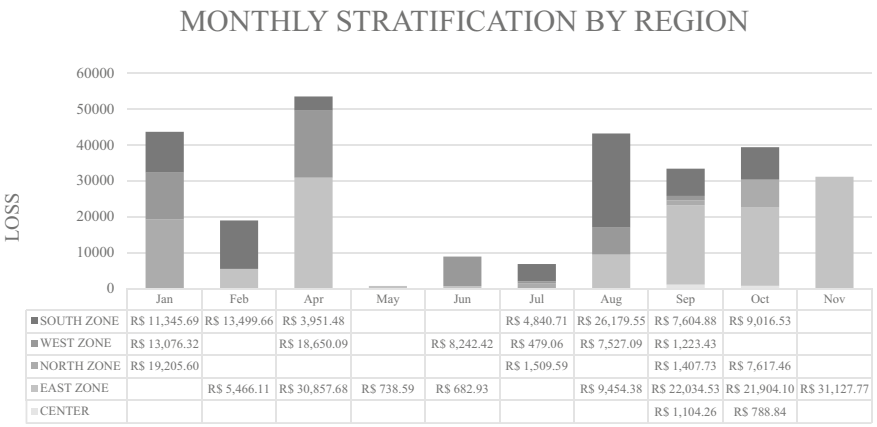
Table 23.1 Number of occurrences of stolen notes/orders in 2023

Date of robbery and theft incidents	Number of notes or orders stolen
	SP-capital
09/January	16
25/January	10
26/January	15
30/January	15
16/February	1
17/February	8
28/February	21
04/April	4
05/April	36
06/April	1
10/April	2
27/April	17
15/May	1
06/June	7
13/June	1
10/July	4
17/July	2
01/August	8
10/August	14
23/August	12
29/August	14
30/August	1
01/September	14
08/September	1
12/September	2
13/September	1
26/September	10
28/September	9
29/September	4
30/September	6
04/October	5
09/October	9
10/October	1
16/October	25
17/October	2
30/October	18
03/November	33
13/November	2
17/November	8
Total cars stolen in 2023: 39	**Total orders stolen in 2023**
	360

Table 23.2 Total losses from theft and robbery in 2023 by region

Total Loss from Robbery and Theft in 2023	CENTER	EAST ZONE	NORTH ZONE	WEST ZONE	SOUTH ZONE	Grand Total
Jan			R$ 19.205,60	R$ 13.076,32	R$ 11.345,69	R$ 43.627,61
Feb		R$ 5.466,11			R$ 13.499,66	R$ 18.965,77
Apr		R$ 30.857,68		R$ 18.650,09	R$ 3.951,48	R$ 53.459,25
May		R$ 738,59				R$ 738,59
Jun		R$ 682,93		R$ 8.242,42		R$ 8.925,35
Jul			R$ 1.509,59	R$ 479,06	R$ 4.840,71	R$ 6.829,36
Aug		R$ 9.454,38		R$ 7.527,09	R$ 26.179,55	R$ 43.161,02
Sep	R$ 1.104,26	R$ 22.034,53	R$ 1.407,73	R$ 1.223,43	R$ 7.604,88	R$ 33.374,83
Oct	R$ 788,84	R$ 21.904,10	R$ 7.617,46		R$ 9.016,53	R$ 39.326,93
Nov		R$ 31.127,77				R$ 31.127,77
Grand Total	**R$ 1.893,10**	**R$ 122.266,09**	**R$ 29.740,38**	**R$ 49.198,41**	**R$ 76.438,50**	**R$ 279.536,48**

Source Prepared by the author based on data provided by the e-commerce company



Graph 23.1 Total losses from theft and robbery in 2023 by region. *Source* Prepared by the author based on data provided by the e-commerce company

management and analysis, as well as risk management, with actions to combat cargo theft being developed to identify the relevant variables employing strategies to reduce thefts.

23.6 Final Considerations

Based on the case study, it was possible to identify the regions and months in which the company suffered the most significant losses from theft and robbery. These results could make it possible to rethink security strategies, such as investment in surveillance technology, IoT systems for tracking goods and escorts for drivers, develop actions to improve last-mile deliveries, and curb and reduce occurrences.

This work makes an academic contribution by addressing this bottleneck in freight transport logistics, as this factor causes much damage to the sector.

Last-mile logistics in areas with the highest crime rates worldwide presents itself as a multifaceted challenge full of unique challenges. In these regions, the delivery of goods faces significant obstacles, from the safety of workers to the integrity of the cargo being transported. Potential threats constantly require innovative logistics strategies beyond simply meeting deadlines. Advanced technologies, such as real-time tracking and dynamic routing algorithms, have emerged as a crucial response to optimize routes, minimize exposure time, and ensure an efficient operation. In addition, close collaboration with local authorities, investments in security, specialized escorts, and proactive prevention measures are essential to deal with the challenges inherent in these complex environments. Ultimately, last-mile logistics in high-crime areas in countries with different realities of local crime rates generally require a holistic approach that combines technological innovation, robust security strategies, and cooperation between the public and private sectors to ensure successful logistics operations.

I would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Private Education Institutions Postgraduate Support Program (PROSUP) for the research incentive grant.

References

1. Ballou, R.H.: Supply Chain Management/Business Logistics, 5th edn., 616 p. Bookman, Porto Alegre (2006). ISBN 978-85-363-0591-2
2. Data Favela.: Pandemic in the Favela: the reality of 14 million favela residents in the fight against the new Coronavirus (2022). https://www.boavistaservicos.com.br/blog/releases/pedidos-de-falencia-caem-160-em2018/. Accessed 13 Jan. 2024
3. Data Favela.: Radiography of the new Brazilian favela. https://entretenimento.band.uol.com.br/. Accessed 14 Jan. 2024; Davis, M.: Planet favela. Bom tempo Editorial (2015)
4. Duarte, A.L.D.C.M., et al.: Last-mile delivery to the bottom of the pyramid in Brazilian slums. Int. J. Phys. Distrib. Logist. Manag. (2019). https://doi.org/10.1108/ijpdlm-01-2018-0008
5. Mendonça, H.G.: E-commerce. Revista IPTEC **4**(2), 240–251 (2016). Accessed 17 Jan. 2024
6. Gevaers, R., Van De Voorde, E., Vanelslander, T.: Characteristics and typology of last-mile logistics from an innovation perspective in an urban context, pp. 56–64. Edward Elgar Publishing, Inc., Cheltenham (2011). https://books.google.com.br/books?hl=enBR&lr=&id=DpYwMe9fBEkC&oi=fnd&pg=PA56&dq=Characteristics+and+typology+of+lastmile+logistics+from+an+innovation+perspective+in+an+urban+context&ots=Gjzj9Y4ZRe&sig=taC2pI9U340fUxVXukSjd8kjeg#v=onepage&q=Characteristics%20and%20typology%20of%20lastmile%20logistics%20from%20an%20innovation%20perspective%20in%20an%20urban%20context&f=false

7. Gil, A.C.: Methods and Techniques of Social Research, 5th edn. Atlas, São Paulo (1999)
8. Laudon, K.C., Laudon, J.P.: Management information systems: managing the digital enterprise. Pearson Prentice Hall, São Paulo (2004)
9. Lim, S.F.W.T., Jin, X., Srai, J.S.: Consumer-driven e-commerce: a literature review, design framework, and research agenda on last-mile logistics models. Int. J. Phys. Distrib. Logist. Manag. **48**(3), 308–332 (2018). https://doi.org/10.1108/IJPDLM-02-2017-0081
10. Maere, B.D.: Ecological and economic impact of automated parcel lockers versus home delivery. Res. Pap. 2016–2017
11. Malhotra, N.K.: Marketing Research: an Applied Orientation. 4th edn. Bookman, Porto Alegre (2004)
12. Manerba.: Use of lockers as an improvement and reduction of last-mile risk in e-commerce. Encontro Nacional de Engenharia de Produção. Santos (2019)
13. Manuj, I., Mentzer, J.T.: Global supply chain risk management. J. Bus. Logist. **29**(1), 133–155 (2008). https://doi.org/10.1002/j.2158-1592.2008
14. Marconi, M.A., Lakatos, E.M.: Methodology of scientific work: basic procedures, bibliographic research, project and report, publications and scientific works. 7. ed. 6.reimpr. Atlas, São Paulo (2011)
15. Meixell, M.J., Norbis, M.: Integrating carrier selection with supplier selection decisions to improve supply chain security. Int. Trans. Oper. Res. **19**(5), 711–732 (2012). https://doi.org/10.1111/j.1475-3995.2011.00817.x
16. Novaes, A.G., Valente, A.M.: Gestão de transportes e frotas, 2nd Revised edn., Cengage Leaming (2008)
17. Oliveira, I.H.I., et al.: Risk management in road freight transport: a study of the paulínia case and fuel transport. ESPACIOS Magazine, vol. 37, no. 03, p. 22 (2016). https://www.revistaespacios.com/a16v37n03/16370322.html. Accessed 6 Jan. 2024
18. Oliveira, R.R.: Cargo theft in Brazil—2017. MC2R—Strategic Intelligence (2018)
19. da Silva, W.M., de Morais, L.A., Frade, C.M., Pessoa, M.F.: Digital marketing, E-commerce and pandemic: a bibliographic review on the Brazilian panorama. Res. Soc. Dev. [S. l.] **10**(5), e45210515054 (2021)
20. Vieira, J.G.V., Fransoo, J.C., Carvalho, C.D.: Freight distribution in megacities: perspectives of shippers, logistics service providers and carriers. J. Transp. Geogr. **46**, 46–54 (2015). https://doi.org/10.1016/j.jtrangeo.2015.05.007
21. Wu, P.J., Chen, M.C., Tsau, C.K.: The data-driven analytics for investigating cargo loss in logistics systems. Int. J. Phys. Distrib. Logist. Manag. **47**(1), 68–83 (2017). https://doi.org/10.1108/ijpdlm-02-2016-0061

Part IV
Statistical Analysis and Model Selection for Complex Structed Data

Chapter 24
Flexible Detection of Birth Cohort Effects on Cancer Mortality

Masayoshi Ishihara, Keisuke Fukui, and Tetsuji Tonda

Abstract Cancer mortality is increasing as the population ages in Japan. Cancer registry data obtained precisely serves as the basis for planning effective cancer control measures. Three time-related factors affect cancer mortality, one of which is the birth cohort effect. Previous descriptive epidemiological studies suggest that the birth cohort effect is not negligible in cancer mortality. In this study, we developed a statistical method to automatically detect and evaluate the statistical significance birth cohort effects in cancer mortality data, using a varying coefficient model. As a result, birth cohort effects detected by the proposed method are in good agreement with previous epidemiological findings.

24.1 Introduction

Although the number of deaths due to cancer is decreasing, it is still the leading cause of death in Japan in 2022. Therefore, it is important to develop an efficient cancer control measures. To this end, identifying accurate trends in cancer risk is necessary. Three time-dependent factors: "age", "period", and "birth cohort" are known to affect cancer mortality. Here, "age" is the age of death, "period" is the year of death, and "birth cohort" is the year of birth. We illustrated effects of these factors using liver cancer mortality data for Japanese males as a typical example in Fig. 24.1. Cancer mortality data are available from the website of the National Cancer Center in Japan [2]. The data are tabulated in 5-year age groups. Figure 24.1 shows the period trends of mortality rates by age groups (left) and the birth cohort trends

M. Ishihara
Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8521, Japan

K. Fukui (✉)
Faculty of Social Safty Sciences, Kansai University, Takatsuki, Osaka 569-1098, Japan
e-mail: kfukui@kansai-u.ac.jp

T. Tonda
Faculty of Regional Development, Prefectural University of Hiroshima, Hiroshima, Hiroshima 734-8558, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_24

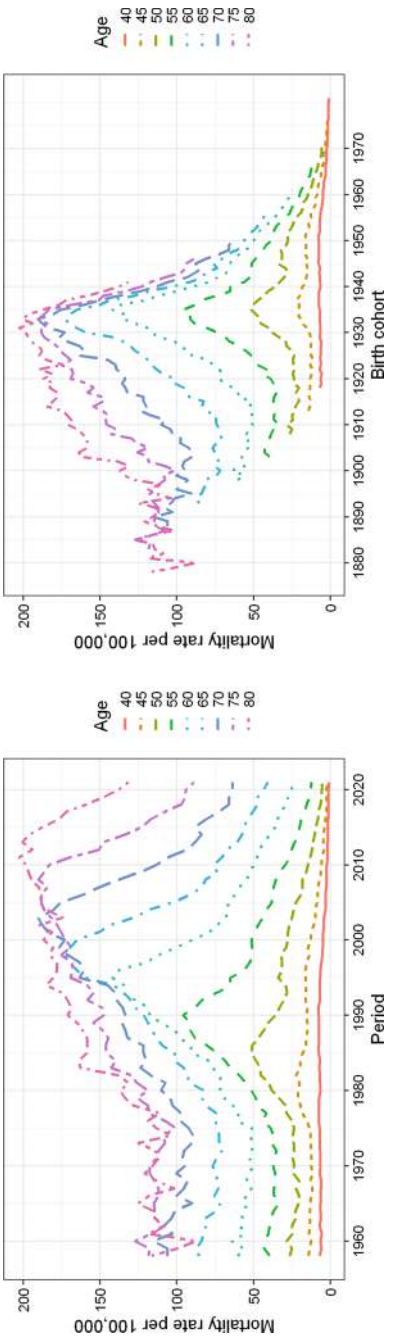


Fig. 24.1 Liver cancer mortality rate by age group for Japanese males

of mortality rates by age groups (right), which allows us to understand the effects of age, period and birth cohort. In fact, The mortality rate of liver cancer increases with age, changes with period and furthermore, a higher mortality rate is seen for a specific birth cohort in Fig. 24.1. Many previous studies have pointed out that the birth cohort born in the 1930s has higher risks of liver cancer, and the reason is thought to be the high infection rate of hepatitis C virus in Japan (see [6, 7, 14, 22]). However, such birth cohort effects are usually not obvious.

Age-Period-Cohort (APC) analysis is a statistical method to evaluate age, period, and birth cohort effects simultaneously (see [13] and [16]). In the APC analysis, a model identification problem arises due to the strict linear dependence among the three variables, $birth\ cohort = period - age$. As a result, it is generally impossible to estimate the three effects separately. Therefore, several approaches have been proposed to solve this problem by estimating the three effects under various assumptions and constraints (see [4, 5, 11, 12]). However, these approaches have often produced inconsistent results.

Kamo et al. in [10] proposed to visualize the mortality risk using mortality data aggregated by age and period. They visualized the risk of cancer mortality in the age-period dimension and searched for birth cohort effects from a geographical point of view. In Kamo et al. in [10], such as birth cohort effects from the interaction of age and period is called the "global" effect. The main purpose of their study was to reveal the trends of cancer risk by visualization, not to determine the existence of birth cohort effects. For this reason, although their method is useful to identify global trends with age and period, it would be difficult to empirically identify birth cohort effects except in extreme cases.

Tonda et al. in [20] developed a statistical method to automatically detect a "local" birth cohort effect, which is not represented in age, period, or global birth cohort effects and evaluate its statistical significance. However, as their method can only detect one local birth cohort effect, it is not possible to detect multiple local birth cohort effects, even when epidemiological findings to suggest them. For example, in the data of liver cancer deaths among Japanese males, Tonda et al. in [20] detected only birth cohort effects around 1930, while Imamura and Sobue in [6] found birth cohort effects around 1940 in addition to those around 1930.

Therefore, we developed a statistical method to automatically detect such non-trivial multiple local birth cohort effects. In this study, we introduced a varying coefficient model and developed a method to simultaneously estimate the varying coefficient and detect multiple local birth cohort effects using a sparse regression approach. This allows us detect local birth cohort effects as well as age and period effects. The proposed method was applied to data on liver and lung cancer mortality rates for Japanese male. Finally, we examined possible reasons for the birth cohort effects detected by the proposed method.

24.2 Method

24.2.1 Varying Coefficient Model

Let $(z_{a,p}, d_{a,p})$ be the population and the observed number of deaths in cancer at age a during period p, respectively, and let $\mathcal{A}$ be the set of the combination of (a, p) (i.e., $(a, p) \in \mathcal{A}$). The number of death in cancer is assumed to follow a Poisson distribution,

$$d_{a,p} \sim \text{Poisson}(z_{a,p}\lambda_{a,p}),\ \log \lambda_{a,p} = \beta_0(a, p),$$

where $\beta_0(a, p)$ is a regression coefficient that varies with (a, p) and $\lambda_{a,p}$ is the mortality rate at (a, p). Regression coefficients that vary with time, geographic location, or other covariates are generally called varying coefficients (see [3]). Here, $\beta_0(a, p)$ represents the risk of death on the age-period plane. Therefore, the age and period specific mortality rates can be calculated as a crude estimate of $\exp(\beta_0(a, p))$.

24.2.2 Estimation of Varying Coeffincients

Tonda et al. in [20] proposed a model of the interaction between age and period. That is, $\beta_0(a, p) = \boldsymbol{\theta}'\boldsymbol{x}(a, p)$. In this study, such a basis vector $\boldsymbol{x}(a, p)$ is used

$$\boldsymbol{x}(a, p) = (1, a, a^2, a^3, a^4, p, p^2, p^3, p^4, ap, ap^2, ap^3, ap^4, a^2p,$$
$$a^2p^2, a^2p^3, a^2p^4, a^3p, a^3p^2, a^3p^3, a^3p^4, a^4p, a^4p^2, a^4p^3, a^4p^4)'.$$

By using the $\boldsymbol{x}(a, p)$, the $\beta_0(a, p)$ represents not only the age and period effects, but also the global trend of the birth cohort effect. Furthermore, Tonda et al. in [20] proposed a new model to detect a local birth cohort effect as follows,

$$\beta_0(a, p) = \boldsymbol{\theta}'\boldsymbol{x}(a, p) + \beta_\mu\phi(\mu, \sigma^2),$$
$$\phi(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p - a - \mu)^2}{2\sigma^2}\right),$$

where μ is mean of normal density term and σ^2 is variance. The method in Tonda et al. [20] assumed only a single local birth cohort effect, while the addition of β_μ allows detection of the local birth cohort effects other than age, period, and global birth cohort effect.

In this study, in order to detect and more flexibly estimate multiple local birth cohort effects, we proposed the new model as follows:

$$\beta_0(a, p) = \boldsymbol{\theta}'\boldsymbol{x}(a, p) + \sum_{\mu \in \mathcal{C}} \beta_\mu\phi(\mu, \sigma^2),$$

where $\mathcal{C}$ is the set of birth cohort years in the data. From the relationship between regression coefficients and relative risk in Poisson regression, the relative risk in each birth year is assumed to be $\exp(\beta_\mu \phi(\mu, \sigma^2))$.

To automatically detect multiple local birth cohort effects that cannot be detected by global birth cohort effects, we added a LASSO penalty term to the following loss function

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \underset{\boldsymbol{\theta}, \boldsymbol{\beta}}{\operatorname{argmin}} \left(-2l(\boldsymbol{\theta}, \boldsymbol{\beta}|\sigma^2) + \lambda \sum_{\mu \in \mathcal{C}} |\beta_\mu| \right),$$

$$l(\boldsymbol{\theta}, \boldsymbol{\beta}|\sigma^2) = \sum_{(a,p) \in \mathcal{A}} \log f(d_{a,p}|z_{a,p}),$$

where $\log f(d_{a,p}|z_{a,p})$ is the logarithm of the probability function of the Poisson distribution expressed by the following equation, $\boldsymbol{\theta}$ is the set of parameters and $\boldsymbol{\beta}$ is the set of coefficients β_μ,

$$\log f(d|z) = d\,(\log z + \beta_0(a, p)) - z e^{\beta_0(a,p)} - \log d!.$$

The parameter estimation and variable selection are performed simultaneously by minimizing the above loss function. The optional tuning parameters λ and σ are obtained by leave-one-out cross-validation method.

24.3 Result

We evaluated the performance of our proposed method for detecting birth cohort effects using liver and lung cancer mortality rates among Japanese males.

24.3.1 Liver Cancer Mortality

We estimated $\beta_0(a, p)$ using a fourth order polynomial interaction basis to automatically detect birth cohort effects. Figure 24.2 shows the relative risk for the birth cohort effect. The maximum relative risk for the birth cohort effect around 1935 was approximately 1.5. In addition, local birth cohort effects were detected between 1920 and 1950. The selected birth cohorts and coefficients are listed in Table 24.1 of Appendix.

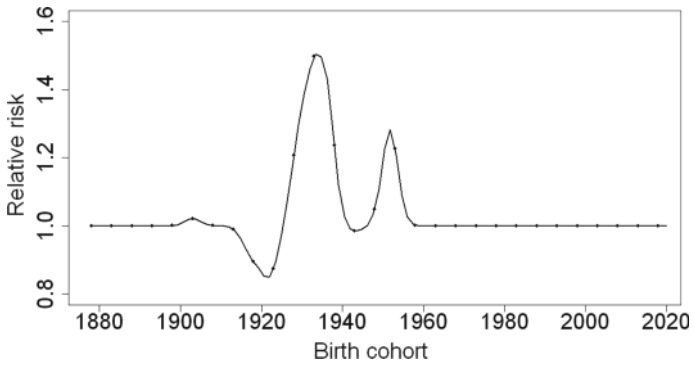


Fig. 24.2 Relative Risk of birth cohort effects on liver cancer mortality in Japanese males

24.3.2 Lung Cancer Mortality

Descriptive epidemiological studies (see [14, 15]) have shown that there was a local peak in lung cancer mortality in the late 1920s and a decreasing trend until the late 1930s. Figure 24.3 shows the trends of lung cancer mortality rates by age group. From these figures, it is difficult to identify the birth cohort effect intuitively unless one is an experienced epidemiologist. Therefore, we applied the proposed method to automatically detect these birth cohort effects. We estimated $\beta_0(a, p)$ using the proposed method. Figure 24.4 shows the relative risk for the birth cohort effect. The minimum relative risk for the birth cohort effect around 1940 is approximately 0.8. Other local birth cohort effects were detected around 1920 and 1950. The selected birth cohorts and coefficients are listed in Table 24.2 of Appendix.

24.4 Discussion

Our proposed method detected birth cohort effects around 1920, 1930, and 1950 for liver cancer mortality. These results are in good agreement with those of previous epidemiological studies (see [6, 7, 10, 21]). In particular, the positive birth cohort effect around 1930 corresponds well to the period after World War II, when intravenous methamphetamine injection and blood trafficking were widespread among the younger generation. The positive birth cohort effect around 1935 is also attributed to the high prevalence of hepatitis C virus infection (see [6, 7, 10, 21]). Considering that Tonda et al. in [20] detected only the cohort around 1935 for the same data, it is considered that the detection is more flexible. As for the lung cancer mortality, birth cohort effects were detected around 1890, 1920, 1940, and 1950. The negative birth cohort effect detected around 1940 is discussed by Marugame et al. in [15]. Although it is well known that smoking is associated with lung cancer, there was an

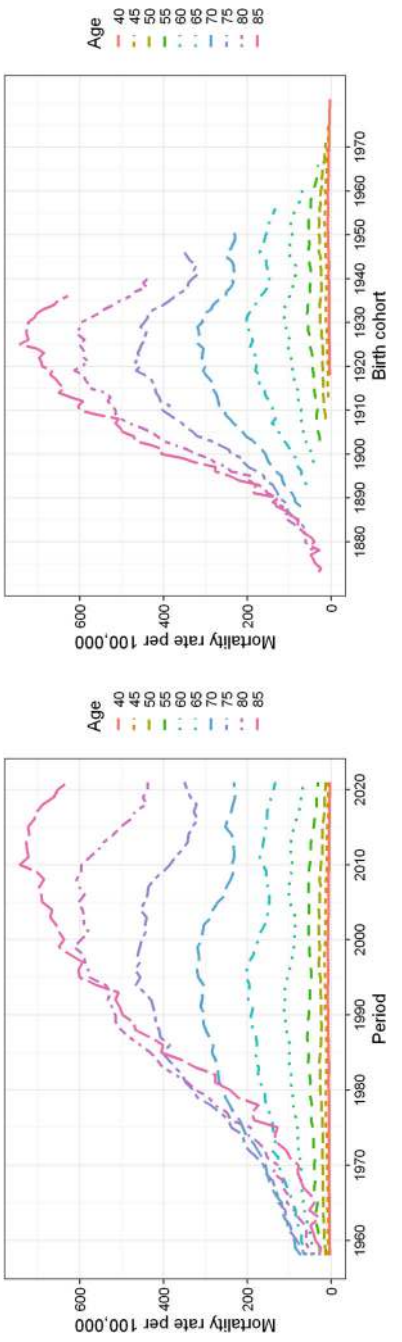


Fig. 24.3 Lung cancer mortality rate by age group for Japanese males

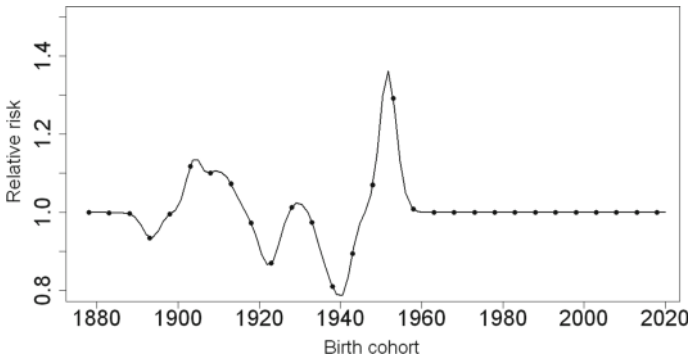


Fig. 24.4 Relative Risk of birth cohort effects on lung cancer mortality in Japanese males

extreme shortage of cigarettes in Japan from the end of World War II to the beginning of Japan's economic growth after World War II. Therefore, males born in the 1930s had less opportunity to start smoking during adolescence. Consequently, the smoking rate of males born in the 1930s was lower than that of females born in the 1930s. The smoking rate among males born in the 1930s declined. Considering the fact that Tonda et al. in [20] detected only the cohorts around 1939 for similar data, it is considered that the proposed method is flexible in detecting birth cohort effects in lung cancer mortality data as well.

In many epidemiological studies in the past, birth cohort effects have been estimated by descriptive assessments of longitudinal behaviors. However, such classical methods are prone to errors due to the influence of researcher's bias and subjectivity on the results. One way to solve this problem is to automatically determine the birth cohort effect using objective statistical methods. For the two cancer types presented in this paper, we have succeeded in automatically detecting the birth cohort effect. However, the proposed method could not cope with overfitting of the model and detected the birth cohorts sensitively, so further improvement is needed in the selection of tuning parameters. This method is expected to be useful for detecting small birth cohort effects, and it could also be useful for developing statistical models for predicting future cancer mortality.

Acknowledgements The authors are grateful to Prof. Wakaki Hirofumi and Asst. Prof. Oda Ryoya for supporting the topic treated in this paper, and this study was supported by JSPS KAKENHI Grant Number 19H01076, 21K17288, and the MHLW Grant Number 23EA0801.

Table 24.1 Birth cohorts detected by the proposed method and their regression coefficients in liver cancer mortality data

Birth cohorts	Coefficients	Birth cohorts	Coefficients
1903	0.0212	1932	0.3816
1916	−0.0668	1933	0.4042
1918	−0.1107	1936	0.3694
1919	−0.1276	1937	0.3047
1920	−0.1469	1942	−0.0090
1921	−0.1639	1943	−0.0146
1922	−0.1630	1951	0.22835
1929	0.2523	1952	0.2464
1930	0.3024		

Appendix

In this section, we report the birth cohorts detected by the proposed method and the estimated regression coefficients. The relative risk for each birth cohort is denoted by $\exp\left(\beta_\mu \phi(\mu, \sigma^2)\right)$.

Liver Cancer Mortality

The tuning parameter in the liver cancer mortality data was $\lambda = 2.8$ and $\sigma = 2.0$. Table 24.1 lists the detected birth cohorts and their regression coefficients.

Lung Cancer Mortality

The tuning parameter in the lung cancer mortality data was $\lambda = 4.6$ and $\sigma = 2.0$. Table 24.2 lists the detected birth cohorts and their regression coefficients.

Table 24.2 Birth cohorts detected by the proposed method and their regression coefficients in lung cancer mortality data

Birth cohorts	Coefficients	Birth cohorts	Coefficients
1893	−0.0670	1924	−0.1136
1894	−0.0646	1929	0.0243
1904	0.1303	1930	0.0245
1905	0.1253	1935	−0.0992
1908	0.0964	1936	−0.1401
1909	0.1008	1938	−0.2090
1910	0.1009	1939	−0.22845
1911	0.0955	1940	−0.2441
1912	0.0859	1941	−0.2247
1913	0.0715	1951	0.2962
1920	−0.0929	1952	0.3064
1921	−0.1257	1954	0.1768
1923	−0.1392	1955	0.1032

References

1. Akita, T., Ohisa, M., Kimura, Y., Fujimoto, M., Miyakawa, Y., Tanaka, J.: Validation and limitation of age-period-cohort model in simulating mortality due to hepatocellular carcinoma from 1940 to 2010 in Japan. Hepatol. Res. **44**, 713–719 (2014). https://doi.org/10.1111/hepr.12177
2. Cancer Statistics. Cancer Information Service, National Cancer Center, Japan (Vital Statistics of Japan, Ministry of Health, Labour and Welfare). https://ganjoho.jp/reg_stat/statistics/data/dl/index.html
3. Hastie, T., Tibshirani, R.: Varying-coefficient models. J. R. Stat. Soc. **55**, 757–779 (1993). https://doi.org/10.1111/j.2517-6161.1993.tb01939.x
4. Holford, T.R.: The estimation of age, period and cohort effects for vital rates. Biometrics 311–324 (1983). https://doi.org/10.22807/2531004
5. Holford, T.R.: Analysing the temporal effects of age, period and cohort. Stat. Methods. Med. Res. 317–337 (1992). https://doi.org/10.1177/096228029200100306
6. Imamura, Y., Sobue, T.: Cancer statistics digest. Mortality trend of colon, rectal, liver, "gallbladder and biliary tract" and pancreas cancer in Japan by birth cohort. Jpn. J. Clin. Oncol. **34**, 491–493 (2004). https://doi.org/10.1093/jjco/hyh085
7. Ishiguro, S., Inoue, M., Tanaka, Y., Mizokami, M., Iwasaki, M., Tsugane, S.: JPHC Study Group: impact of viral load of hepatitis C on the incidence of hepatocellular carcinoma: a population-based cohort study (JPHC Study). Cancer Lett. **300**, 173–179 (2011). https://doi.org/10.1016/j.canlet.2010.10.002
8. Ito, Y., Ioka, A., Nakayama, T., Tsukuma, H., Nakamura, T.: Comparison of trends in cancer incidence and mortality in Osaka, Japan, using an age-period-cohort model. Asian Pac. J. Cancer Prev. **12**, 879–888 (2011). https://journal.waocp.org/article_25626.html
9. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer, New York (2013). https://link.springer.com/content/pdf/10.1007/978-1-0716-1418-1.pdf
10. Kamo, K.: Cancer mortality risk visualization on age-period plane. Proc. Inst. Stat. Math. **59**, 217 (2011). https://ndlsearch.ndl.go.jp/books/R000000004-I023488856

11. Keyes, K.M., Utz, R.L., Robinson, W., Li, G.: What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. Soc. Sci. Med. **70**, 1100–1108 (2010). https://doi.org/10.1016/j.socscimed.2009.12.018
12. Kupper, L.L., Janis, J.M., Karmous, A., Greenberg, B.G.: Statistical age-period-cohort analysis: a review and critique. J. Chronic. Dis. **38**, 811–830 (1985). https://doi.org/10.1016/0021-9681(85)90105-5
13. Mason, K.O., Mason, W.M., Winsborough, H.H., Poole, W.K.: Some methodological issues in cohort analysis of archival data. Am. Sociol. Rev. 242–258 (1973). https://doi.org/10.22807/2094398
14. Matsuda, T., Saika, K.: Trends in liver cancer mortality rates in Japan, USA, UK, France and Korea based on the WHO mortality database. Jpn. J. Clin. Oncol. **42**, 360–361 (2012). https://doi.org/10.1093/jjco/hys048
15. Marugame, T., Kamo, K., Sobue, T., Akiba, S., Mizuno, S., Satoh, H., Suzuki, T., Tajima, K., Tamakoshi, A., Tsugane, S.: Trends in smoking by birth cohorts born between 1900 and 1977 in Japan. Prev. Med. **42**, 120–127 (2006). https://doi.org/10.1016/j.ypmed.2005.09.009
16. Frenk, S.M., Yang, Y.C., Land, K.C.: Assessing the significance of cohort and period effects in hierarchical age-period-cohort models: Applications to verbal test scores and Voter Turnout in U.S. Presidential Elections. Soc. Forces **92**, 221–248 (2013). https://doi.org/10.1093/sf/sot066
17. Takahashi, H., Okada, M., Kano, K.: Age-period-cohort analysis of lung cancer mortality in Japan, 1960–1995. J. Epidemiol. **11**, 151–159 (2001). https://doi.org/10.2188/jea.11.151
18. Tanaka, H., Uera, F., Tsukuma, H., Ioka, A., Oshima, A.: Distinctive change in male liver cancer incidence rate between the 1970s and 1990s in Japan: comparison with Japanese-Americans and US whites. Jpn. J. Clin. Oncol. **37**, 193–196 (2007). https://doi.org/10.1093/jjco/hyl148
19. Tango, T.: Estimation of age, period and cohort effects: decomposition into linear trend and curvature components. Jpn. J. Appl. Stat. 14, 45–49 (in Japanese) (1985). https://cir.nii.ac.jp/crid/1573105974896221696
20. Tonda, T., Satoh, K., Kamo, K.: Detecting a local cohort effect for cancer mortality data using a varying coefficient model. J. Epidemiol. **25**, 639–646 (2015). https://doi.org/10.2188/jea.JE20140218
21. Tonda, T., Satoh, K., Nakayama, T., Katanoda, K., Sobue, T., Ohtaki, M.: A nonparametric mixed-effects model for cancer mortality. Aust. N. Z. J. Stat. **53**, 247–256 (2011). https://doi.org/10.1111/j.1467-842X.2011.00615.x
22. Yoshimi, I., Sobue, T.: Mortality trend of liver cancer in Japan 1960–2000. Jpn. J. Clin. Oncol. **33**, 202–203 (2003). https://europepmc.org/article/MED/12816081

Chapter 25
Non-parametric Bias-Reduction Estimation of Residual Variance in Varying Coefficient Regression Model

Hirokazu Yanagihara and Sanai Shibayama

Abstract We propose a non-parametric higher order asymptotic unbiased estimator of residual variance in the varying coefficient regression model with q smoothing variables. The estimation method is based on a local linear fitting and does not require optimization of hyperparameters such as the bandwidth of the kernel function or the smoothing parameters of the penalized smoothing spline. The order of the bias of the proposed estimator is $O(n^{-4/q})$ under appropriate conditions.

25.1 Introduction

The present paper is concerned with the following varying coefficient regression model:

$$y_i = \beta_0(\boldsymbol{z}_i) + \sum_{j=1}^{k} x_{ij}\beta_j(\boldsymbol{z}_i) + \varepsilon_i, \ \ (i = 1, \ldots, n), \tag{25.1}$$

where n is the sample size, y_i is a response variable, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ are k non-stochastic explanatory variables, $\beta_0(\boldsymbol{z}), \beta_1(\boldsymbol{z}), \ldots, \beta_k(\boldsymbol{z})$ are k unknown coefficients varying with respect to $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^q$ ($\mathcal{Z}$ is a closed domain), $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})'$ are q non-stochastic smoothing variables (the term "smoothing variable" follows [8]), and ε_i is an error variable. Here, we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed with mean 0 and unknown residual variance σ^2, and the prime $'$ denotes the transpose of a matrix or vector.

The varying coefficient regression model with a single smoothing variable was proposed by [5]. Thereafter, many authors have proposed varying coefficient regression models with multiple smoothing variables (for details, see the review paper [8]). A local liner estimation using kernel functions or a penalized smoothing spline using basis functions is used to estimate the varying coefficients (see, e.g., [8]).

H. Yanagihara (✉) · S. Shibayama
Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8526, Japan
e-mail: yanagi-hiro@hiroshima-u.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_25

The bandwidth of the kernel functions and the smoothing parameter of the penalized smoothing spline are commonly selected by minimizing model selection criteria such as the C_p criterion proposed by [6]. It is well known that the C_p criterion requires an asymptotic unbiased estimator of σ^2. It would be absurd if the model selection criteria for selecting a bandwidth or a smoothing parameter depended on an optimized it. Therefore, it is important to provide an asymptotic unbiased estimator of σ^2 that is independent of the optimization of the bandwidth or smoothing parameter. Such an estimator would not only allow us to propose a new C_p criterion for the choice of bandwidth or smoothing parameter, as in [9], but we could also use it as a threshold to check overfitting, as in [10].

In the present paper, we extend the non-parametric estimation proposed in [3] to construct an asymptotic unbiased estimator of σ^2. Their estimation method is based on a local linear fitting (LLF) and is only applicable when there is only β_0 with a single smoothing variable z. In the method, using two individuals a_1 and a_2 which have distinct smoothing variables z_ℓ ($\ell \in \{a_1, a_2\}$) that are close to z_i, we estimate σ^2 from the residual between y_i and a straight line through two data pairs $\{(z_\ell, y_\ell) \mid \ell \in \{a_1, a_2\}\}$. This estimator has been shown to be a higher-order bias-corrected estimator, i.e., the order of the bias is $O(n^{-4})$, under appropriate conditions, if we assume that the maximum distance to the nearest smoothing variable is $O(n^{-1})$. The estimator was extended to allow multiple smoothing variables by [1, 7]. In our proposed method, using $(q+1)$ individuals $a_1, ..., a_{q+1}$ which have distinct vectors of smoothing variables $\boldsymbol{z}_\ell$ ($\ell \in \{a_1, \ldots, a_{q+1}\}$) that are close to $\boldsymbol{z}_i$, we non-parametrically construct an estimate of a varying coefficient by a hyperplane through $(q+1)$ data pairs $\{(\boldsymbol{z}'_\ell, y_\ell) \mid \ell \in \{a_1, \ldots, a_{q+1}\}\}$. Then, since there are k_1 $(= k+1)$ varying coefficients in the model (25.1), including the intercept, the residuals are constructed using p data pairs $\{(\boldsymbol{z}'_\ell, y_\ell) \mid \ell \in \{a_1, \ldots, a_p\}\}$, where $p = (q+1)k_1$. Our proposed estimator is also a higher-order bias-corrected estimator. Roughly speaking, if the maximum $\|\boldsymbol{z}_i - \boldsymbol{z}_\ell\|$ over $\ell \in \{a_1, \ldots, a_p\}$ is assumed to be $O(n^{-1/q})$, then the bias of the new estimator of σ^2 is $O(n^{-4/q})$.

The remainder of the paper is organized as follows. In Sect. 25.2, we propose our higher-order asymptotic unbiased estimator of σ^2. In Sect. 25.3, we present a simple numerical study to verify that the proposed estimator properly reduces the bias. Technical details are provided in the Appendix.

25.2 New Estimator Based on the LLF

We first define a set of integers that show an individual having a vector of smoothing variables that is close to the vector of the ith individual. Let $d_{ij} = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|$ and $\delta_{i(1)}, \ldots, \delta_{i(n)}$ be the values of $d_{i1}, \ldots, d_{in}$ rearranged in decreasing order. Then, we define the following set of integers:

$$\mathcal{S}_i(m) = \left\{ \ell \in \{1, \ldots, n\} \backslash \{i\} \;\middle|\; d_{ij} \leq \delta_{i(m+1)} \right\}. \tag{25.2}$$

It should be kept in mind that $\#(\mathcal{S}_i(m)) \geq m$ holds because there may be overlapping in $z_1, \ldots, z_n$.

Let $\boldsymbol{V}_i(m)$ and $\boldsymbol{V}_{0,i}(m)$ be $p \times p$ ($p = (q+1)(k+1)$) and $k_1 \times k_1$ ($k_1 = k+1$) matrices as

$$\boldsymbol{V}_i(m) = \sum_{\ell \in \mathcal{S}_i(m)} \boldsymbol{u}_\ell \boldsymbol{u}_\ell', \quad \boldsymbol{V}_{0,i}(m) = \sum_{\ell \in \mathcal{S}_i(m)} \boldsymbol{u}_{0,\ell} \boldsymbol{u}_{0,\ell}', \tag{25.3}$$

where $\boldsymbol{u}_\ell$ and $\boldsymbol{u}_{0,\ell}$ are the p-dimensional and k_1-dimensional vectors defined by

$$\boldsymbol{u}_\ell = \begin{pmatrix} 1 \\ \boldsymbol{z}_\ell \end{pmatrix} \otimes \boldsymbol{u}_{0,\ell}, \quad \boldsymbol{u}_{0,\ell} = \begin{pmatrix} 1 \\ \boldsymbol{x}_\ell \end{pmatrix}. \tag{25.4}$$

Here the notation $\otimes$ denotes the Kronecker product (see, e.g., Chap. 16 in [4]). We assume here that at least the rank of $\boldsymbol{W} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is p. In addition, the following p-dimensional and k_1-dimensional vectors are also defined:

$$\boldsymbol{h}_i(m) = \sum_{\ell \in \mathcal{S}_i(m)} y_\ell \boldsymbol{u}_\ell, \quad \boldsymbol{h}_{0,i}(m) = \sum_{\ell \in \mathcal{S}_i(m)} y_\ell \boldsymbol{u}_{0,\ell}. \tag{25.5}$$

We will now use the individuals belonging to $\mathcal{S}_i(p)$ to construct hyperplanes, which are LLFs of the varying coefficients, i.e., $\beta_j(\boldsymbol{z}) \approx \boldsymbol{\eta}_j' \boldsymbol{z}$. However, to uniquely determine the hyperplanes, $\boldsymbol{V}_i(p)$ is required to be of full rank. Hence, $\mathcal{S}_i(p)$ has to be modified to $\mathcal{S}_i(p + a_i)$ to satisfy the full-rank condition, where a_i is the integer defined by

$$a_i = \min \left\{ a \in [n_{p+1}] \,|\, \mathrm{rank}(\boldsymbol{V}_i(p+a)) = p \right\}. \tag{25.6}$$

Here $n_m = n - m$ and $[m]$ is defined by $[m] = \{0, 1, \ldots, m\}$. It should be noted that the definition of set $[m]$ usually does not include 0 as an element, but here it does. Of course, in many cases, a_i becomes 0 when there is no overlapping in $z_1, \ldots, z_n$. Also, the sample points $\{z_\ell \mid \ell \in \mathcal{S}_i(p + a_i)\}$ must have at least $(q+1)$ different vectors as elements.

If there are k_1 or more individuals with the same vector of smoothing variables z_i, i.e., $^{\exists}\ell \in \{k_1, \ldots, n-1\}$ $s.t.$ $\delta_{i(\ell+1)} = 0$, then it is better to estimate the varying coefficients of the ith individual by ordinary multiple regression using those individuals. However, to uniquely determine regression coefficients, $\boldsymbol{V}_{0,i}(p)$ is required to be of full rank. If a set $\mathcal{B}_i$, described below, is not the empty set, then $\mathcal{S}_i(k_1)$ is modified to $\mathcal{S}_i(k_1 + b_i)$ to satisfy the full-rank condition, where b_i is the integer defined by

$$b_i = \min(\mathcal{B}_i), \ \mathcal{B}_i = \left\{ b \in [n_{k+2}] \,\middle|\, \mathrm{rank}(\boldsymbol{V}_{0,i}(k_1 + b)) = k_1, \delta_{i(k_1+1+b)} = 0 \right\}. \tag{25.7}$$

Using $\mathcal{S}_i(m)$ with a_i in (25.6) or b_i in (25.7), the following estimator of σ^2 based on a new estimator of the residual of an individual is proposed:

$$\hat{\sigma}^2_{\mathrm{LLF}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}_i^2,\ \tilde{\varepsilon}_i = \begin{cases} \dfrac{y_i - \boldsymbol{u}_i'\boldsymbol{V}_i(p+a_i)^{-1}\boldsymbol{h}_i(p+a_i)}{\sqrt{1+\boldsymbol{u}_i'\boldsymbol{V}_i(p+a_i)^{-1}\boldsymbol{u}_i}} & (\mathcal{B}_i = \emptyset) \\ \dfrac{y_i - \boldsymbol{u}_{0,i}'\boldsymbol{V}_{0,i}(k_1+b_i)^{-1}\boldsymbol{h}_{0,i}(k_1+b_i)}{\sqrt{1+\boldsymbol{u}_{0,i}'\boldsymbol{V}_{0,i}(k_1+b_i)^{-1}\boldsymbol{u}_{0,i}}} & (\mathcal{B}_i \neq \emptyset) \end{cases}, \tag{25.8}$$

where $\boldsymbol{V}_i(m)$ and $\boldsymbol{V}_{0,i}(m)$ are the $p \times p$ and $k_1 \times k_1$ matrices given in (25.3), and $\boldsymbol{h}_i(m)$ and $\boldsymbol{h}_{0,i}(m)$ are the p- and k_1-dimensional vectors given in (25.5).

To evaluate the order of the bias of $\hat{\sigma}^2_{\mathrm{LLF}}$ relative to σ^2, the following four conditions are assumed.

A1 $\delta_{\mathrm{M}} = \max_{i=1,\ldots,n} \delta_{i(p+a_i+1)} = O(n^{-1/q})$ as $n \to \infty$.

A2 All varying coefficients $\beta_0(\boldsymbol{z}), \beta_1(\boldsymbol{z}), \ldots, \beta_k(\boldsymbol{z})$ are C^2 functions of $\boldsymbol{z} \in \mathcal{Z}$.

A3 $f_{\mathrm{M}} = \max_{i=1,\ldots,n} \#(\mathcal{F}_i(p)) = O(1)$ as $n \to \infty$, where $\mathcal{F}_i(p)$ is defined by $\mathcal{F}_i(p) = \{\ell \in \{1,\ldots,n\} | \{i\} \cap \mathcal{S}_\ell(p+a_\ell)\}$.

A4 $\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X})/n = O(1)$ as $n \to \infty$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$.

When $q = 1$, it is natural to assume that the maximum difference in the nearest smoothing variable is $O(n^{-1})$, as assumed in [3]. When $q = 2$, if $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are placed on a lattice, then the numbers of data placed on the vertical and horizontal axes are $\sqrt{n}$. Without loss of generality, it should be clear that the distance between adjacent lattices is $1/\sqrt{n}$ if the lattices are arranged in a square of length 1. This fact illustrates the propriety of assumption A1. Assumption A3 specifies how the number of sample points increases as n increases. Since $\#(\mathcal{F}_i(p))$ is the frequency with which the ith individual appears in sets $\mathcal{S}_1(p+a_1)$ to $\mathcal{S}_n(p+a_n)$, assumption A3 is an acceptable condition. Assumption A4 is a condition often used to evaluate asymptotic behavior in multiple linear regression. When assumptions A1, A2, A3, and A4 hold, $E[\hat{\sigma}^2_{\mathrm{LLF}}] = \sigma^2 + O(n^{-4/q})$ (the proof is given in the Appendix).

25.3 Numerical Study

In this section, a simple numerical study is performed to verify that the proposed estimator properly reduces the bias for σ^2. In particular, the numerical study considers the coordinates of the location as $\boldsymbol{z}$ at $q = 2$. This model is used in the so-called geographically weighted regression (GWR) model proposed by [2].

We generated simulation data using the model in (25.1) with $k = 5$ and $n = 30, 50, 100, 300$, and 500. Error variables were generated independently from the normal distribution with mean 0 and variance $\sigma^2 = 0.25$ and $\boldsymbol{X} = \boldsymbol{X}_0(0.7 \times \boldsymbol{I}_5 + 0.3 \times \boldsymbol{1}_5\boldsymbol{1}_5')^{1/2}$, where $\boldsymbol{X}_0$ is an $n \times 5$ matrix with all elements generated independently from the uniform distribution over $(-1, 1)$ and $\boldsymbol{1}_k$ is the k-dimensional vector of ones.

Smoothing variables were generated independently from the uniform distribution over $(0, 3)$. Since we allow overlap in $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, we examined three patterns in the

numerical study: Pattern 1 with intense overlap (different sample points are 30% of n), Pattern 2 with mild overlap (different sample points are 80% of n), and Pattern 3 with no overlap. Figure 25.1 shows placements of sample points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. In the figure, the red indicates the points of intense overlap. In addition, the first, second, and last columns show the placements in Patterns 1, 2, and 3, respectively, and the first through fifth rows show the placements for $n = 30, 50, 100, 300$, and 500, respectively.

The following functions with contour lines as shown in Fig. 25.2 were used as the true varying coefficients.

- $\beta_0(\boldsymbol{z}) = 0.27725 \times z_1 - 0.08650 \times z_2 + 0.90718$.
- $\beta_1(\boldsymbol{z}) = \exp[-\{(z_1 - 0.00052)^2 + (z_2 - 0.28998)^2\}/2]$.
- $\beta_2(\boldsymbol{z}) = \sin(0.73399 \times z_1)\exp\{-(z_1 - 0.51061)^2\} - \exp\{-(z_2 - 0.88070)^2\}$.
- $\beta_3(\boldsymbol{z}) = \exp[-\{(z_1 - 0.01405)^2 + (z_2 - 0.12322)^2\}/2]$.
- $\beta_4(\boldsymbol{z}) = -0.48177 \times \{(z_1 - 0.06469)^2 + (z_2 - 0.17511)^2\}$.
- $\beta_5(\boldsymbol{z}) = \sin(0.33061 \times z_1)\exp\{-(z_1 - 0.95485)^2\} - \exp\{-(z_2 - 0.44075)^2\}$.

As a competitor to the proposed estimator, we prepared the estimator of residual variance when the model was estimated by GWR. The Gaussian kernel with bandwidth ν was used as the geographically weighted functions. Let $\boldsymbol{H}_\nu$ be an $n \times n$ hat matrix from the GWR model. Then, the estimator of σ^2 derived from the GWR model is given by $\hat{\sigma}^2_\nu = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H}_\nu)'(\boldsymbol{I} - \boldsymbol{H}_\nu)\boldsymbol{y}/\mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{H}_\nu)'(\boldsymbol{I} - \boldsymbol{H}_\nu)\}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)'$. In the GWR, we used $\nu = \nu_0, \nu_0/2, \nu_0/4$, and $\nu_0/7$ as the bandwidths, where ν_0 is the minimum radius of the circle such that it is centered at $\boldsymbol{z}_i$ and contains k_1 or more sample points including $\boldsymbol{z}_i$, i.e., $\nu_0 = \max_{i=1,\ldots,n} \delta_{i(k_1)}$. The values used for ν_0 are listed in Table 25.1.

Table 25.2 lists the percentage bias of the estimator relative to the true value (denoted by Bias) and the mean squared error of the estimator standardized by the squared true value (denoted by MSE), which are defined by

$$\text{Bias} : 100 \times \left(\frac{E[\hat{\sigma}^2] - \sigma^2}{\sigma^2} \right), \quad \text{MSE} : E\left[\left(\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2} \right)^2 \right],$$

where $\hat{\sigma}^2$ is an estimator of σ^2. The estimators used are $\hat{\sigma}_{\mathrm{LLF}}$ in (25.8) (denoted by Proposed) and $\hat{\sigma}^2_\nu$ with $\nu = \nu_0, \nu_0/2, \nu_0/4$, and $\nu_0/7$ (denoted by $\nu = \nu_0$, $\nu = \nu_0/2$, $\nu = \nu_0/4$, and $\nu = \nu_0/7$). From the numerical results, we can see that $\hat{\sigma}^2_{\mathrm{LLF}}$ most successfully reduces the bias. The estimator $\hat{\sigma}^2_\nu$ does not provide effective bias reduction without optimizing the bandwidths. There were $\hat{\sigma}^2_\nu$ with $\nu = \nu_0/2$ and $\nu_0/4$ for which MSE was lower than that of $\hat{\sigma}^2_{\mathrm{LLF}}$ when n was large. We think this is because $Var[\hat{\sigma}^2_{\mathrm{LLF}}]$ tends to become large relative to $Var[\hat{\sigma}^2_\nu]$, and the rate of decrease of $Var[\hat{\sigma}^2_\nu]$ with respect to n is large. Although the MSE of $\hat{\sigma}^2_{\mathrm{LLF}}$ is not the smallest, the difference from the smallest value is not large, and so does not call into question the performance of $\hat{\sigma}^2_{\mathrm{LLF}}$ when n is large.

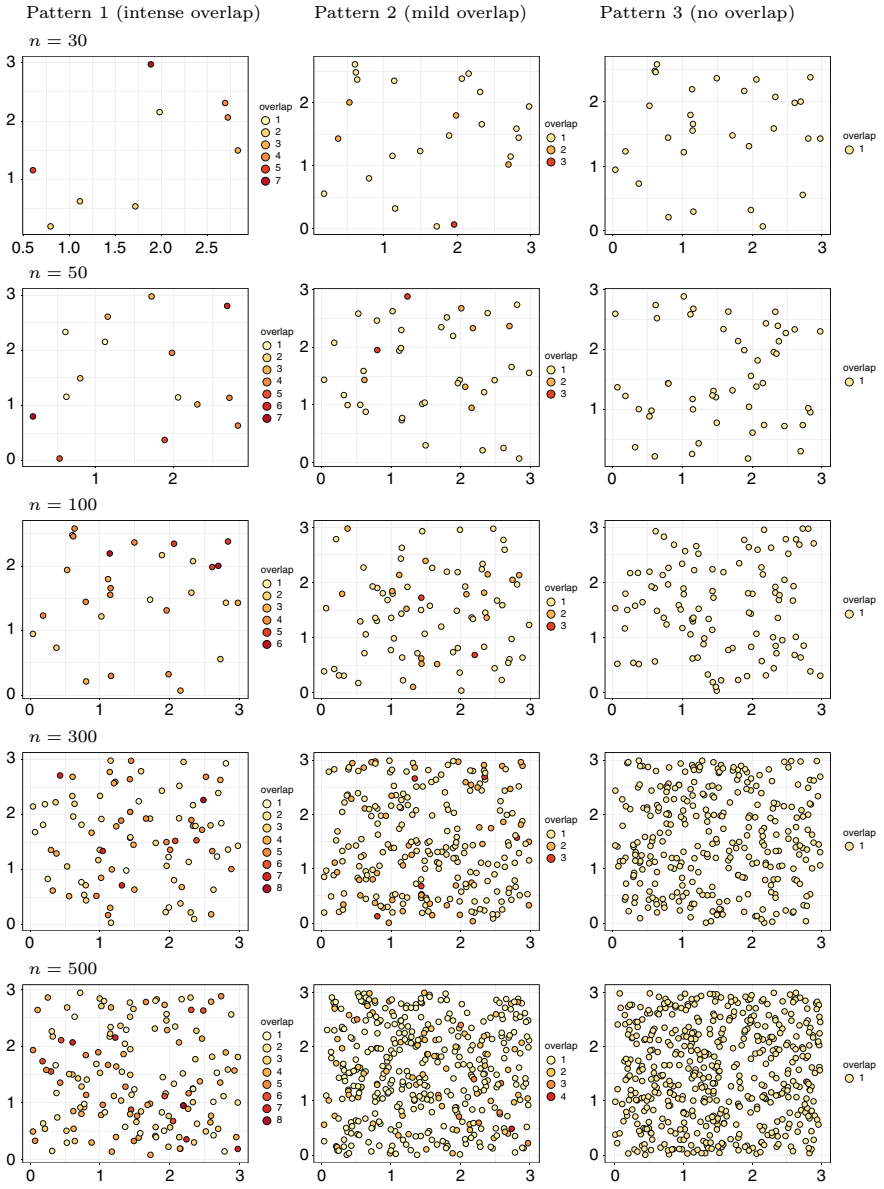


Fig. 25.1 Placements of sample points $z_1, \ldots, z_n$

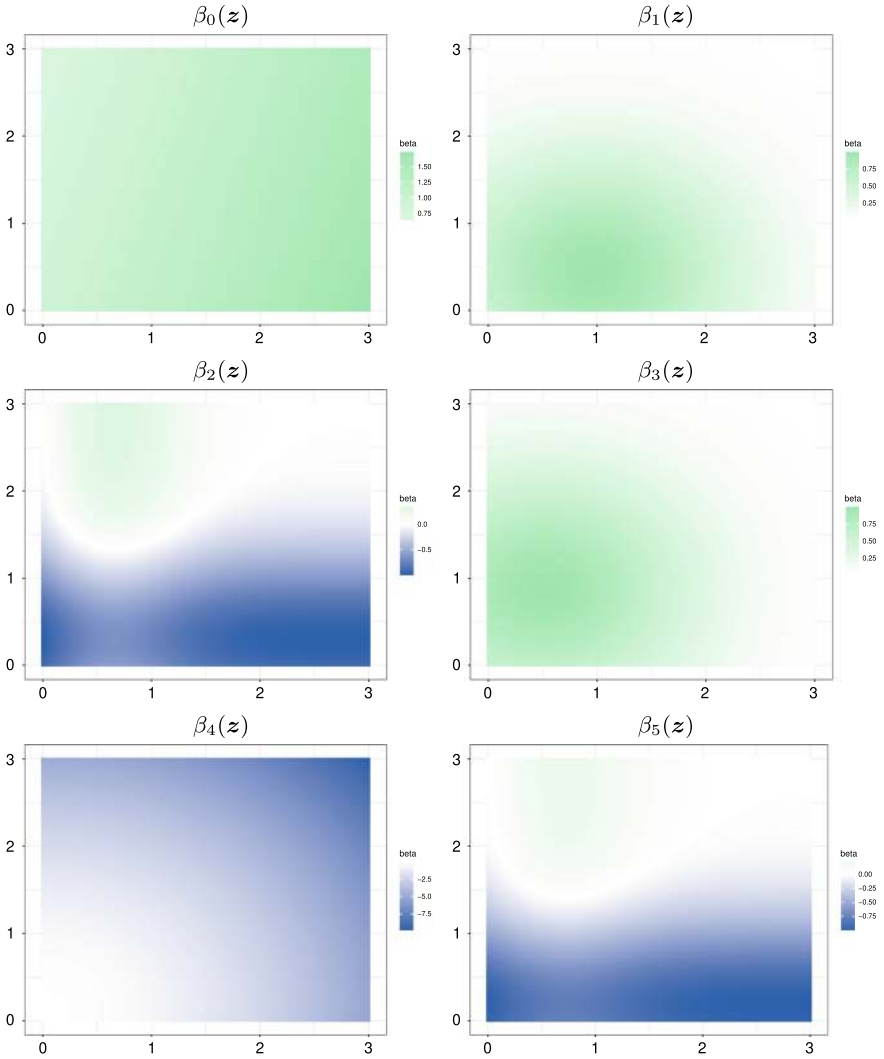


Fig. 25.2 Contour lines for true varying coefficients

Table 25.1 Values of ν_0

	$n = 30$	$n = 50$	$n = 100$	$n = 300$	$n = 500$
Pattern 1	0.9842	0.9228	0.7785	0.4623	0.4331
Pattern 2	1.2051	1.2551	0.9930	0.4261	0.3246
Pattern 3	1.3586	1.1213	0.9862	0.4014	0.3933

Table 25.2 Biases and MSEs of the proposed estimator and competitor estimators

n	Proposed		GWR							
			($\nu = \nu_0$)		($\nu = \nu_0/2$)		($\nu = \nu_0/4$)		($\nu = \nu_0/7$)	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Pattern 1 (with intensity overlap)										
30	6.79	0.50	99.58	1.18	21.41	0.32	12.18	0.89	7.34	1.13
50	9.43	0.24	182.38	3.48	40.95	0.34	11.71	0.80	1.02	2.24
100	4.86	0.11	123.68	1.58	37.82	0.19	13.02	0.10	6.82	0.16
300	−0.04	0.04	44.22	0.21	11.13	0.02	3.26	0.02	0.14	0.06
500	0.60	0.02	34.83	0.13	7.83	0.01	2.87	0.01	1.73	0.03
Pattern 2 (with mild overlap)										
30	10.29	0.48	257.21	6.99	80.26	0.96	51.57	1.27	59.22	4.02
50	3.07	0.24	344.67	12.10	88.83	0.90	38.58	0.38	24.91	0.62
100	1.22	0.12	285.09	8.23	84.76	0.77	29.24	0.17	10.86	0.27
300	−0.08	0.03	27.32	0.08	12.14	0.03	5.10	0.04	2.17	0.25
500	0.52	0.02	17.02	0.03	6.40	0.01	3.62	0.03	1.58	0.14
Pattern 3 (with no overlap)										
30	7.92	0.65	285.78	8.60	127.30	1.93	31.36	0.97	15.53	2.32
50	7.02	0.29	276.10	7.82	84.40	0.85	32.51	0.40	20.09	0.97
100	1.40	0.12	290.98	8.54	62.26	0.43	26.09	0.15	8.57	0.25
300	0.00	0.03	28.06	0.09	12.49	0.03	3.97	0.06	0.68	0.35
500	0.48	0.02	27.72	0.08	8.34	0.01	4.54	0.02	2.18	0.13

Acknowledgements The authors wish to thank two reviewers for their helpful comments. This research was supported by JSPS Bilateral Program Grant Number JPJSBP 120219927 and JSPS KAKENHI Grant Numbers 20H04151, 23H00809, and the ISM Specially Promoted Research (2023-ISMCRP-4106). The authors also thank FORTE Science Communications (https://www.forte-science.co.jp/) for English language editing.

25.4 Appendix: Proof of the Order of the Bias

Obviously, $E[\tilde{\varepsilon}_i^2] = \sigma^2$ when $\mathcal{B}_i \neq \emptyset$, so we consider the case when $\mathcal{B}_i = \emptyset$, where $\tilde{\varepsilon}_i$ and $\mathcal{B}_i$ are given by (25.8) and (25.7), respectively. Let $\boldsymbol{g}_j(\boldsymbol{z})$ be the q-dimensional gradient vector of $\beta_j(\boldsymbol{z})$ and $\boldsymbol{H}_j(\boldsymbol{z})$ be the $q \times q$ Hessian matrix of $\beta_j(\boldsymbol{z})$, which are given by

$$\boldsymbol{g}_j(\boldsymbol{z}) = \frac{\partial}{\partial \boldsymbol{z}}\beta_j(\boldsymbol{z}), \quad \boldsymbol{H}_j(\boldsymbol{z}) = \frac{\partial^2}{\partial \boldsymbol{z}\partial \boldsymbol{z}'}\beta_j(\boldsymbol{z}).$$

By applying the Taylor expansion to $\beta_j(\boldsymbol{z}_\ell)$ ($\ell \in \mathcal{S}_i(p + a_i)$) at $\boldsymbol{z}_i$, we have

$$\beta_j(\boldsymbol{z}_\ell) = \beta_j(\boldsymbol{z}_i) + \boldsymbol{g}_j(\boldsymbol{z}_i)'(\boldsymbol{z}_\ell - \boldsymbol{z}_i) + \xi_{j,i,\ell}, \tag{25.9}$$

where $\mathcal{S}_i(m)$ is the set of integers given by (25.2) and $\xi_{j,i,\ell}$ is the remainder term of the Taylor expansion, given by $\xi_{j,i,\ell} = (\boldsymbol{z}_\ell - \boldsymbol{z}_i)'\boldsymbol{H}_j(\boldsymbol{\zeta}_{j,i,\ell})(\boldsymbol{z}_\ell - \boldsymbol{z}_i)/2$. Here, $\boldsymbol{\zeta}_{j,i,\ell}$ is some real vector between $\boldsymbol{z}_i$ and $\boldsymbol{z}_\ell$. It follows from assumption A2 that $|\xi_{j,i,\ell}| \leq \tau_j d_{i\ell}^2/2$, where $d_{i\ell} = \|\boldsymbol{z}_i - \boldsymbol{z}_\ell\|$. This is because $\beta_j(\boldsymbol{z})$ is a C^2 function on the closed domain $\mathcal{Z}$. This implies that the minimum and maximum eigenvalues of $\boldsymbol{H}_j(\boldsymbol{z})$ are bounded. Hence, it is sufficient to use the larger absolute value of the minimum and maximum eigenvalues of $\boldsymbol{H}_j(\boldsymbol{z})$ as τ_j. Let $\boldsymbol{\beta}(\boldsymbol{z}) = (\beta_0(\boldsymbol{z}), \beta_1(\boldsymbol{z}), \ldots, \beta_k(\boldsymbol{z}))'$ be a k_1-dimensional vector of varying coefficients, and $\mu_\ell(\boldsymbol{z}) = \boldsymbol{\beta}(\boldsymbol{z})'\boldsymbol{u}_{0,\ell}$, where $\boldsymbol{u}_{0,\ell}$ is the k_1-dimensional vector given in (25.4). Using the expansion in (25.9), we have

$$\mu_\ell(\boldsymbol{z}_\ell) = \mu_\ell(\boldsymbol{z}_i) + \{(\boldsymbol{z}_\ell - \boldsymbol{z}_i) \otimes \boldsymbol{u}_{0,\ell}\}'\boldsymbol{\theta}_i + \alpha_{i,\ell},$$

where $\boldsymbol{\theta}_i$ is a $k_1 q$-dimensional vector and $\alpha_{i,\ell}$ is a scalar, given by

$$\boldsymbol{\theta}_i = \mathrm{vec}\left(\left(\boldsymbol{g}_0(\boldsymbol{z}_i), \boldsymbol{g}_1(\boldsymbol{z}_i), \ldots, \boldsymbol{g}_k(\boldsymbol{z}_i)\right)'\right), \quad \alpha_{i,\ell} = \boldsymbol{\xi}_{i,\ell}'\boldsymbol{u}_{0,\ell},$$

Here $\boldsymbol{\xi}_{i,\ell} = (\xi_{0,i,\ell}, \xi_{1,i,\ell}, \ldots, \xi_{k,i,\ell})'$ is a k_1-dimensional vector and $\mathrm{vec}(\boldsymbol{A})$ denotes an operator that transforms a matrix $\boldsymbol{A}$ to a vector by stacking the first to last columns of $\boldsymbol{A}$ (see, e.g., chap. 16 in [4]). Let $p_i = \#(\mathcal{S}_i(p + a_i))$, $\boldsymbol{y}_i$ be a p_i-dimensional vector of sequentially stacked y_ℓ ($\ell \in \mathcal{S}_i(p + a_i)$), $\boldsymbol{U}_{0,i}$ be the $p_i \times k_1$ matrix of sequentially stacked $\boldsymbol{u}_{0,\ell}'$ ($\ell \in \mathcal{S}_i(p + a_i)$), and $\boldsymbol{W}_i$ be the $p_i \times p$ matrix of sequentially stacked $\boldsymbol{u}_\ell'$ ($\ell \in \mathcal{S}_i(p + a_i)$). Then, $\tilde{\varepsilon}_i$ can be rewritten as

$$\tilde{\varepsilon}_i = \frac{y_i - \boldsymbol{u}_i'(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\boldsymbol{W}_i'\boldsymbol{y}_i}{\sqrt{1 + \boldsymbol{u}_i'(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\boldsymbol{u}_i}} = \frac{y_i - \boldsymbol{c}_i'\boldsymbol{y}_i}{\sqrt{1 + \|\boldsymbol{c}_i\|^2}}, \tag{25.10}$$

where $\boldsymbol{c}_i$ is the p_i-dimensional vector given by $\boldsymbol{c}_i = \boldsymbol{W}_i(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\boldsymbol{u}_i$. Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be the $p \times k_1$ and $p \times k_1 q$ matrices defined by

$$\boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{I}_{k_1} \\ \boldsymbol{O}_{k_1 q, k_1} \end{pmatrix}, \quad \boldsymbol{D}_2 = \begin{pmatrix} \boldsymbol{O}_{k_1, k_1 q} \\ \boldsymbol{I}_{k_1 q} \end{pmatrix},$$

where $\boldsymbol{O}_{m,q}$ is the $m \times q$ matrix of zeros. Then, it is easy to see that $\boldsymbol{W}_i \boldsymbol{D}_1 = \boldsymbol{U}_{0,i}$. Let $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\alpha}_i$ be the p_i-dimensional vectors of sequentially stacked ε_ℓ and $\alpha_{i,\ell}$ ($\ell \in \mathcal{S}_i(p + a_i)$), respectively. It follows from these equations that

$$\boldsymbol{y}_i = \boldsymbol{W}_i \boldsymbol{D}_1 \boldsymbol{\beta}(z_i) + \boldsymbol{W}_i \boldsymbol{D}_2 \boldsymbol{\theta}_i - \left(\boldsymbol{z}_i' \otimes \boldsymbol{U}_{0,i}\right) \boldsymbol{\theta}_i + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i.$$

Notice that

$$\begin{aligned}
\boldsymbol{c}_i' \boldsymbol{W}_i \boldsymbol{D}_1 \boldsymbol{\beta}(z_i) &= \boldsymbol{u}_{0,1}' \boldsymbol{\beta}(z_i) = \mu_i(z_i), \ \boldsymbol{c}_i' \boldsymbol{W}_i \boldsymbol{D}_2 \boldsymbol{\theta}_i = \left(z_i \otimes \boldsymbol{u}_{0,i}\right)' \boldsymbol{\theta}_i, \\
\boldsymbol{c}_i' \left(\boldsymbol{z}_i' \otimes \boldsymbol{U}_{0,i}\right) \boldsymbol{\theta}_i &= \boldsymbol{u}_i' (\boldsymbol{W}_i' \boldsymbol{W}_i)^{-1} \boldsymbol{W}_i' \left(z_{i1} \boldsymbol{W}_i \boldsymbol{D}_1, \ldots, z_{iq} \boldsymbol{W}_i \boldsymbol{D}_1\right) \boldsymbol{\theta}_i \\
&= \left(z_i \otimes \boldsymbol{u}_{0,i}\right)' \boldsymbol{\theta}_i.
\end{aligned}$$

These imply that

$$\boldsymbol{c}_i' \boldsymbol{y}_i = \mu_i(z_i) + \boldsymbol{c}_i' (\boldsymbol{\varepsilon}_i + \boldsymbol{\alpha}_i). \tag{25.11}$$

Substituting the result $y_i = \mu_i(z_i) + \varepsilon_i$ and (25.11) into (25.10) yields

$$\tilde{\varepsilon}_i = \frac{\varepsilon_i - \boldsymbol{c}_i' \boldsymbol{\varepsilon}_i}{\sqrt{1 + \|\boldsymbol{c}_i\|^2}} - \frac{\boldsymbol{c}_i' \boldsymbol{\alpha}_i}{\sqrt{1 + \|\boldsymbol{c}_i\|^2}} = r_i - \gamma_i.$$

Notice that $E[r_i] = 0$ and

$$E\left[r_i^2\right] = \frac{1}{1 + \|\boldsymbol{c}_i\|^2} \left\{E\left[\varepsilon_i^2\right] + E\left[\boldsymbol{c}_i' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \boldsymbol{c}_i\right]\right\} = \frac{1}{1 + \|\boldsymbol{c}_i\|^2} \left(1 + \|\boldsymbol{c}_i\|^2\right) \sigma^2 = \sigma^2.$$

Hence, we have $E[\hat{\sigma}^2_{\mathrm{LLF}}] = \sigma^2 + n^{-1} \sum_{i=1}^n \gamma_i^2$. Let $\tau_{\mathrm{M}} = \max\{\tau_0, \tau_1, \ldots, \tau_k\}$. Recall that $|\xi_{j,i,\ell}| \leq \tau_j d_{i\ell}^2 / 2$. It follows from the Cauchy-Schwarz inequality that

$$\gamma_i^2 \leq \frac{\|\boldsymbol{c}_i\|^2 \|\boldsymbol{\alpha}_i\|^2}{1 + \|\boldsymbol{c}_i\|^2} < \|\boldsymbol{\alpha}_i\|^2 = \sum_{\ell \in \mathcal{S}_i(p + a_i)} \left(\boldsymbol{\xi}_{i,\ell}' \boldsymbol{u}_{0,\ell}\right)^2 \leq \frac{1}{4} k_1 \tau_{\mathrm{M}}^2 \delta_{\mathrm{M}}^4 \sum_{\ell \in \mathcal{S}_i(p + a_i)} \|\boldsymbol{u}_\ell\|^2.$$

This implies that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \gamma_i^2 &\leq \frac{1}{4} k_1 \tau_{\mathrm{M}}^2 \delta_{\mathrm{M}}^4 \left\{\frac{1}{n} \sum_{i=1}^n \sum_{\ell \in \mathcal{S}_i(p + a_i)} \left(1 + \|\boldsymbol{x}_\ell\|^2\right)\right\} \\
&= \frac{1}{4} k_1 \tau_{\mathrm{M}}^2 \delta_{\mathrm{M}}^4 \left\{\frac{1}{n} \sum_{i=1}^n \#(\mathcal{F}_i(p)) \left(1 + \|\boldsymbol{x}_i\|^2\right)\right\} \\
&\leq \frac{1}{4} k_1 \tau_{\mathrm{M}}^2 \delta_{\mathrm{M}}^4 f_{\mathrm{M}} \left\{1 + \frac{1}{n} \mathrm{tr}\left(\boldsymbol{X}' \boldsymbol{X}\right)\right\} = O(n^{-4/q}).
\end{aligned}$$

Consequently, $E[\hat{\sigma}^2_{\mathrm{LLF}}] = \sigma^2 + O(n^{-4/q})$ is proved.

References

1. Bock, M., Bowman, A.W., Ismail, B.: Estimation and inference for error variance in bivariate nonparametric regression. Stat. Comput. **17**, 39–47 (2007). https://doi.org/10.1007/s11222-006-9000-0
2. Brunsdon, C., Fotheringham, S., Charlton, M.: Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. **28**, 281–298 (1996). https://doi.org/10.1111/j.1538-4632.1996.tb00936.x
3. Gasser, T., Sroka, L., Jennen-Steinmetz, C.: Residual variance and residual pattern in nonlinear regression model. Biometrika **73**, 625–633 (1986). https://doi.org/10.2307/2336527
4. Harville, D.A.: Matrix Algebra from a Statistician's Perspective. Springer, New York (1997)
5. Hastie, T., Tibshirani, R.: Varying-coefficient models. J. Roy. Stat. Soc. Ser. B **55**, 757–796 (1993). https://doi.org/10.1111/j.2517-6161.1993.tb01939.x
6. Mallows, C.L.: Some comments on C_p. Technometrics **15**, 661–675 (1973). https://doi.org/10.2307/1267380
7. Ohtaki, M.: Some estimators of covariance matrix in multivariate nonparametric regression and their applications. Hiroshima Math. J. **20**, 63–91 (1990). https://doi.org/10.32917/hmj/1206454441
8. Park, B.U., Mammen, E., Lee, Y.K., Lee, E.R.: Varying coefficient regression models: a review and new developments. Int. Stat. Rev. **83**, 36–64 (2015). https://doi.org/10.1111/insr.12029
9. Yanagihara, H.: A non-iterative optimization method for smoothness in penalized spline regression. Stat. Comput. **22**, 527–544 (2012). https://doi.org/10.1007/s11222-011-9245-0
10. Yanagihara, H., Ohtaki, M.: On avoidance of the over-fitting in the B-spline non-parametric regression model. Jpn. J. Appl. Stat. **33**, 51–69 (2004) (in Japanese). https://doi.org/10.5023/jappstat.33.51

Chapter 26
Coordinate Descent Algorithm of the Group Lasso for Selecting Between-Individual Explanatory Variables in the Three-Mode GMANOVA Model

Rei Monden, Keito Horikawa, Isamu Nagai, and Hirokazu Yanagihara

Abstract Data consisting of three different entities (e.g., individual, item, and time) are commonly referred to as three-mode data. The present study focused on three-mode data where one entity is time, namely longitudinal multivariate data collected from the same set of individuals. To identify time trends underlying the three-mode data, a three-mode GMANOVA model was proposed. This model expresses three-mode data by means of a matrix containing explanatory variables for individuals, for items, for a function of time trend and parameters. Although algorithms for estimating the three-mode GMANOVA model have been proposed, the available algorithms require a predefined matrix for explanatory variables to differentiate individuals. However, selecting proper explanatory variables prior to the analysis can be challenging in practice. Thus, the present study proposes an algorithm to select explanatory variables for individuals by means of the group Lasso.

26.1 Introduction

Data on different entities (e.g., individual, item, and time) are commonly referred to as *three-mode data*. The current study focuses on three-mode data for which one entity represents time. This setup is common for longitudinal multivariate data collected from a common set of units (typically individuals). One simple example consists of the collection of height and weight data from a set of individuals through the same

R. Monden (✉) · K. Horikawa
Informatics and Data Science Program, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8521, Japan
e-mail: mondenr@hiroshima-u.ac.jp

I. Nagai
Faculty of Liberal Arts and Sciences, Chukyo University, Nagoya, Aichi 466-8666, Japan

H. Yanagihara
Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8526, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_26

span of years. For a single item, the generalized multivariate analysis of variance (GMANOVA) model can be applied to estimate a longitudinal time trend [5].

The GMANOVA model was recently extended to handle two or more items or variables simultaneously [3], as follows. Consider a set of n individuals that responded to m items at p time points. Then for each item d ($d = 1, \ldots, m$), there is an $n \times p$ matrix $\boldsymbol{Y}_d = (\boldsymbol{y}_{d1}, \ldots, \boldsymbol{y}_{dn})'$ containing the responses from the n individuals collected at the p time points. Upon concatenating column-wise the m data matrices $\boldsymbol{Y}_d$ for $d = 1, \ldots, m$, we obtain the $n \times mp$ matrix $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$. The three-mode GMANOVA model proposed by [3] can then be written

$$\boldsymbol{Y} \sim \mathcal{N}_{n \times mp}\left(\mathbf{1}_n \boldsymbol{\mu}'(\boldsymbol{C} \otimes \boldsymbol{X})' + \boldsymbol{A}\Theta(\boldsymbol{C} \otimes \boldsymbol{X})', \Psi \otimes \Sigma \otimes \boldsymbol{I}_n\right), \tag{26.1}$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones, $\boldsymbol{\mu}$ is an lq-dimensional unknown vector of intercepts, $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)'$ is an $n \times k$ between-individuals matrix containing the values of the n individuals on k explanatory variables, assumed to have centered columns (i.e., $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$, where $\mathbf{0}_k$ is the k-dimensional vector of zeros) and $\text{rank}(\boldsymbol{A}) = k$, $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$ denotes a $k \times lq$ matrix of unknown coefficients for the longitudinal time trend, $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)'$ is an $m \times l$ between-items matrix containing the scores of the m items on l explanatory variables about each item with $\text{rank}(\boldsymbol{C}) = l$, $\boldsymbol{X}$ is a $p \times q$ matrix which captures the longitudinal time trend with $\text{rank}(\boldsymbol{X}) = q$, and Ψ and Σ are $m \times m$ and $p \times p$ unknown positive definite matrices, respectively. The $(1, 1)$th element of Ψ is fixed to be 1 to ensure that Ψ and Σ can be estimated. Moreover, $\otimes$ denotes the Kronecker product (see, e.g., Chap. 16 in [2]), and the notation $\boldsymbol{Y} \sim \mathcal{N}_{n \times mp}(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{S}})$ denotes that an $n \times mp$ random matrix $\boldsymbol{Y}$ follows the matrix normal distribution with the $n \times mp$ mean matrix $\boldsymbol{\mathcal{M}}$ ($= E[\boldsymbol{Y}]$) and the $nmp \times nmp$ covariance matrix $\boldsymbol{\mathcal{S}}$ ($= Cov[\text{vec}(\boldsymbol{Y})]$), where $\text{vec}(\boldsymbol{Y})$ denotes an operator that transforms a matrix $\boldsymbol{Y}$ to a vector by stacking the columns of $\boldsymbol{Y}$ from first to last (see, e.g., Chap. 16 in [2]). An example of applying three-mode GMANOVA model to real data can be found in [3].

For the three-mode GMANOVA model to capture longitudinal time trends, it is important that appropriate explanatory variables for the between-individuals matrix $\boldsymbol{A}$ to be selected. The selection of these explanatory variables is increasingly challenging given the abundance of data that can now be stored and potentially analyzed. It is therefore necessary to consider variable selection when there are many available between-individual explanatory variables. Such a variable selection method has not yet been considered for the three-mode GMANOVA model. To fill this gap in the literature, in the present paper, we adapt the group Lasso regression proposed by Yuan and Lin [8] to the three-mode GMANOVA model. In order to also satisfy the oracle property [6], we incorporate the adaptive Lasso weight proposed by Zou [10] as a penalty term in the three-mode GMANOVA model. However, introducing this penalty term makes it impossible to obtain a closed-form solution from joint minimization of the penalized negative log-likelihood function. As a workaround, we use the coordinate descent algorithm (CDA) to optimize the regression coefficients. Our contribution is based on extending the results of Yanagihara and Oda [7] to update the CDA for the current GMANOVA setting. We further apply an existing update

algorithm from the block-wise CDA [3] to estimate the remaining parameters other than the regression coefficients.

This paper is organized as follows. In Sect. 26.2, we introduce the three-mode GMANOVA model, the group Lasso approach, and associated algorithm. In Sect. 26.3, we perform a series of simple numerical experiments to evaluate whether our proposed model can indeed select relevant between-individual explanatory variables. In Sect. 26.4, we summarize the main findings from our study and suggest possible avenues for future research.

26.2 Group Lasso for the Three-Mode GMANOVA Model and Its Algorithm

In order to select columns in $\boldsymbol{A}$ based on the group Lasso, we assume $\boldsymbol{A}$ to have columns of unit length, i.e., the diagonal elements of $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{A}$ are 1. Here, we express $\boldsymbol{A}'\boldsymbol{A}$ as $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. For estimating $\boldsymbol{\mu}$, Θ, Ψ, and Σ, the negative log-likelihood function is defined as below [3]:

$$\ell_0(\boldsymbol{\mu}, \Theta, \Psi, \Sigma) = \frac{1}{2}\left\{nmp \log 2\pi + n \log\left(|\Psi|^p |\Sigma|^m\right) + \Delta_{\Psi\otimes\Sigma}(\boldsymbol{Y}, \Gamma)\right\}, \quad (26.2)$$

where $\Gamma = (\boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\Theta)(\boldsymbol{C} \otimes \boldsymbol{X})'$, and $\Delta_S(\boldsymbol{B}_1, \boldsymbol{B}_2)$ is the sum of the squared Mahalanobis distances between two $n \times mp$ matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$: $\Delta_S(\boldsymbol{B}_1, \boldsymbol{B}_2) = \mathrm{tr}\{(\boldsymbol{B}_1 - \boldsymbol{B}_2)\boldsymbol{S}^{-1}(\boldsymbol{B}_1 - \boldsymbol{B}_2)'\}$. We need to add a group Lasso-type penalty term for selecting the columns of $\boldsymbol{A}$. Thus, we define the objective function as follows:

$$\ell_\lambda(\boldsymbol{\mu}, \Theta, \Psi, \Sigma) = \ell_0(\boldsymbol{\mu}, \Theta, \Psi, \Sigma) + \lambda \sum_{j=1}^{k} \omega_j \|\boldsymbol{\theta}_j\|, \quad (26.3)$$

where ω_i is the weight for the adaptive Lasso proposed by [10]. The weight ω_j is usually set as the inverse of the Euclidean norm of the jth row vector of the maximum likelihood estimate for Θ (i.e., the group Lasso estimate with $\lambda = 0$). In this paper, we also use this setting.

As described in a previous study [3], we can estimate Ψ and Σ as below

$$\hat{\Psi} = \frac{1}{np}\sum_{i=1}^{n} \hat{\boldsymbol{\mathcal{E}}}_i \hat{\Sigma}^{-1} \hat{\boldsymbol{\mathcal{E}}}_i', \quad \hat{\Sigma} = \frac{1}{nm}\sum_{i=1}^{n} \hat{\boldsymbol{\mathcal{E}}}_i' \hat{\Psi}^{-1} \hat{\boldsymbol{\mathcal{E}}}_i, \quad (26.4)$$

where $\hat{\boldsymbol{\mathcal{E}}}_i$ is the $m \times p$ matrix defined by

$$\hat{\boldsymbol{\mathcal{E}}}_i = \left(\boldsymbol{y}_{1i} - (\boldsymbol{c}_1' \otimes \boldsymbol{X})(\hat{\boldsymbol{\mu}} + \hat{\Theta}'\boldsymbol{a}_i), \ldots, \boldsymbol{y}_{mi} - (\boldsymbol{c}_m' \otimes \boldsymbol{X})(\hat{\boldsymbol{\mu}} + \hat{\Theta}'\boldsymbol{a}_i)\right)',$$

in which $\hat{\boldsymbol{\mu}}$ and $\hat{\Theta}$ are estimates of $\boldsymbol{\mu}$ and Θ, to be defined later.

If we fix Ψ and Σ, the objective function (26.3) can be further split into two parts: one part dependent on and one part independent from $\boldsymbol{\mu}$ and Θ. The former part is the penalized multivariate residual sum of squares, which can be written as shown below:

$$\mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta) = \frac{1}{2}\Delta_{\Psi\otimes\Sigma}(\boldsymbol{Y}, \Gamma) + \lambda \sum_{j=1}^{k} \omega_j \|\boldsymbol{\theta}_j\|. \tag{26.5}$$

However, unlike in [3], the estimator for Θ cannot be obtained in an explicit form. As a workaround, we apply the CDA to estimate Θ given $\boldsymbol{\mu}$, Ψ, and Σ.

First, we split the objective function (26.5) into three terms: a constant term, a term that depends on $\boldsymbol{\mu}$, and a term that depends on Θ with the penalty part:

$$\mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta) = \frac{1}{2}\mathrm{tr}\{\boldsymbol{Y}(\Psi^{-1} \otimes \Sigma^{-1})\boldsymbol{Y}'\} + h(\boldsymbol{\mu}) + f_\lambda(\Theta), \tag{26.6}$$

where $h(\boldsymbol{\mu}) = n\boldsymbol{\mu}'(\boldsymbol{C}'\Psi^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})\boldsymbol{\mu} - \mathbf{1}_n'\boldsymbol{Y}(\Psi^{-1}\boldsymbol{C} \otimes \Sigma^{-1}\boldsymbol{X})\boldsymbol{\mu}/2$ and

$$\begin{aligned} f_\lambda(\Theta) = {} & \frac{1}{2}\mathrm{tr}\{\boldsymbol{V}\Theta(\boldsymbol{C}'\Psi^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})\Theta'\} \\ & - \mathrm{tr}\{\Theta(\boldsymbol{C}'\Psi^{-1} \otimes \boldsymbol{X}'\Sigma^{-1})\boldsymbol{Y}'\boldsymbol{A}\} + \lambda \sum_{j=1}^{k} \omega_j \|\boldsymbol{\theta}_j\|. \end{aligned}$$

Given the estimates $\hat{\Psi}$ and $\hat{\Sigma}$, the estimator for $\boldsymbol{\mu}$ can be found similarly to as in the ordinary GMANOVA model (i.e., the root of the equation $\partial h(\boldsymbol{\mu})/\partial\boldsymbol{\mu} = \mathbf{0}_{lq}$ after substituting $\hat{\Psi}$ and $\hat{\Sigma}$ into Ψ and Σ, respectively):

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}(\boldsymbol{C}'\hat{\Psi}^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{C}'\hat{\Psi}^{-1} \otimes \boldsymbol{X}'\hat{\Sigma}^{-1})\boldsymbol{Y}'\mathbf{1}_n. \tag{26.7}$$

We still need to determine the estimate for Θ given $\hat{\Psi}$ and $\hat{\Sigma}$. In this case, since $\arg\min_\Theta \mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta) = \arg\min_\Theta f_\lambda(\Theta)$, we focus on obtaining the partial derivative of $f_\lambda(\Theta)$ with respect to each $\boldsymbol{\theta}_j$ $(j = 1, \ldots, k)$. In order to simplify expressions, we define $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k) = \boldsymbol{Y}'\boldsymbol{A}$, where $\boldsymbol{z}_j$ $(j = 1, \ldots, k)$ is an mp-dimensional vector. Then, from $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$, we can rewrite $f_\lambda(\Theta)$ as $f_\lambda(\Theta) = f_0(\Theta) + \lambda \sum_{j=1}^{k} \omega_j \|\boldsymbol{\theta}_j\|$, where

$$f_0(\Theta) = \frac{1}{2}\sum_{a,b=1}^{k} v_{ab}\boldsymbol{\theta}_a'(\boldsymbol{C}'\Psi^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})\boldsymbol{\theta}_b - \sum_{j=1}^{k} \boldsymbol{\theta}_j'(\boldsymbol{C}'\Psi^{-1} \otimes \boldsymbol{X}'\Sigma^{-1})\boldsymbol{z}_j,$$

and v_{ab} is the (a, b)th element of $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{A}$, and $\sum_{a,b=1}^{k}$ denotes $\sum_{a=1}^{k}\sum_{b=1}^{k}$.

26.2.1 The $\hat{\boldsymbol{\theta}}_r = \mathbf{0}_{lq}$ Case

Firstly, we identify the conditions for the estimate of $\boldsymbol{\theta}_r$ to be equal to $\mathbf{0}_{lq}$, for $r = 1, \ldots, k$. To this end, we consider the gradient of the tangent to $f_\lambda(\Theta)$ at $\boldsymbol{\theta}_r = \mathbf{0}_{lq}$ while keeping the other $\boldsymbol{\theta}_j$ $(j \neq r)$ fixed. However, the penalty term in $f_\lambda(\Theta)$ is not differentiable at $\boldsymbol{\theta}_r = \mathbf{0}_{lq}$. To overcome this issue, we consider $\boldsymbol{\theta}_r = \eta\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^{lq} \mid \|\boldsymbol{\alpha}\| = 1\}$, differentiate with respect to η, and then consider $\eta \to +0$.

Substituting $\eta\boldsymbol{\alpha}$ into $\boldsymbol{\theta}_r$ in the penalty term $\lambda \sum_{j=1}^k \omega_j \|\boldsymbol{\theta}_j\|$ and ignoring all $\boldsymbol{\theta}_j$ terms with $j \neq r$, we obtain $\lambda\omega_r|\eta|$, since $\|\boldsymbol{\alpha}\| = 1$. Then, differentiating with respect to η, the penalty term becomes $\lambda\omega_r \operatorname{sign}(\eta)$, where $\operatorname{sign}(\delta)$ is 1 for $\delta > 0$, -1 for $\delta < 0$, and 0 for $\delta = 0$.

Here, we consider the derivative of $f_0(\Theta)$ with respect to $\boldsymbol{\theta}_r$ for each $r = 1, \ldots, k$. Let $\Theta_{[-r]}$ be the $k \times lq$ matrix with the rth row replaced by $\mathbf{0}_{lq}$, i.e., $\Theta_{[-r]} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{r-1}, \mathbf{0}_{lq}, \boldsymbol{\theta}_{r+1}, \ldots, \boldsymbol{\theta}_k)'$. It should be emphasized that $\Theta_{[-r]}$ does not depend on $\boldsymbol{\theta}_r$ at all. Defining $\boldsymbol{g}_r(\boldsymbol{\theta}_r) = -\partial f_0(\Theta)/\partial\boldsymbol{\theta}_r$, we obtain

$$\boldsymbol{g}_r(\boldsymbol{\theta}_r) = (\boldsymbol{C}'\Psi^{-1} \otimes \boldsymbol{X}'\Sigma^{-1})\boldsymbol{z}_r - (\boldsymbol{C}'\Psi^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})\left(\Theta'_{[-r]}\boldsymbol{v}_r + \boldsymbol{\theta}_r\right),$$

since we assume $v_{rr} = 1$ for $r = 1, \ldots, k$, and $\Theta'_{[-r]}\boldsymbol{v}_r = \sum_{j\neq r}^k v_{jr}\boldsymbol{\theta}_j$. Thus, using the chain rule, we have $\partial f_0(\Theta_{[-r]} + \eta\boldsymbol{e}_r\boldsymbol{\alpha}')/\partial\eta = -\boldsymbol{g}_r(\eta\boldsymbol{\alpha})'\boldsymbol{\alpha}$ for $\eta \neq 0$, where $\boldsymbol{e}_r$ is the k-dimensional unit vector whose rth element is 1 and other elements are 0. Moreover, we observe that

$$\boldsymbol{g}_r(\mathbf{0}_{lq}) = \left(\boldsymbol{C}'\Psi^{-1} \otimes \boldsymbol{X}'\Sigma^{-1}\right)\left\{\boldsymbol{z}_r - (\boldsymbol{C} \otimes \boldsymbol{X})\Theta'_{[-r]}\boldsymbol{v}_r\right\}, \tag{26.8}$$

which is equal to $\lim_{\eta\to+0} \boldsymbol{g}_r(\eta\boldsymbol{\alpha})$. The previous expression is the main term of the gradient of the tangent to $f_0(\Theta)$ at $\boldsymbol{\theta}_r = \mathbf{0}_{lq}$ with the other $\boldsymbol{\theta}_j$ $(j \neq r)$ fixed.

We note that the $\boldsymbol{\theta}_r$ that minimizes $f_\lambda(\Theta)$ is zero when

$$\forall\boldsymbol{\alpha} \in \mathcal{A},\ -\boldsymbol{g}(\mathbf{0}_{lq})'\boldsymbol{\alpha} + \lambda\omega_r \geq 0. \tag{26.9}$$

Here, using the Schwarz inequality, the sufficient condition for (26.9) is $\|\boldsymbol{g}_r(\mathbf{0}_{lq})\| \leq \lambda\omega_r$. Hence, the sufficient condition for $\hat{\boldsymbol{\theta}}_r = \arg\min_{\boldsymbol{\theta}_r} f_\lambda(\Theta)$ to be $\mathbf{0}_{lq}$ for given estimates $\hat{\Psi}$, $\hat{\Sigma}$, and $\hat{\boldsymbol{\theta}}_j$ $(j \neq r)$ is

$$\left\|\left(\boldsymbol{C}'\hat{\Psi}^{-1} \otimes \boldsymbol{X}'\hat{\Sigma}^{-1}\right)\left\{\boldsymbol{z}_r - (\boldsymbol{C} \otimes \boldsymbol{X})\hat{\Theta}'_{[-r]}\boldsymbol{v}_r\right\}\right\| \leq \lambda\omega_r, \tag{26.10}$$

where $\hat{\Theta}_{[-r]}$ is the $k \times lq$ matrix defined by replacing $\boldsymbol{\theta}_j$ in $\Theta_{[-r]}$ with $\hat{\boldsymbol{\theta}}_j$ $(j \neq r)$. Thus, if the above condition is satisfied at some r $(r = 1, \ldots, k)$, we set $\hat{\boldsymbol{\theta}}_r = \mathbf{0}_{lq}$, for given $\hat{\Psi}$, $\hat{\Sigma}$ and $\hat{\boldsymbol{\theta}}_j$ $(j \neq r)$.

26.2.2 The $\hat{\boldsymbol{\theta}}_r \neq \mathbf{0}_{lq}$ Case

When $\hat{\boldsymbol{\theta}}_r \neq \mathbf{0}_{lq}$, $\hat{\boldsymbol{\theta}}_r$ is obtained as the vector satisfying

$$\frac{\partial}{\partial \boldsymbol{\theta}_r} \mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta)\Big|_{\boldsymbol{\theta}_r = \hat{\boldsymbol{\theta}}_r} = \mathbf{0}_{lq}.$$

Recall that we can now rewrite (26.6) as

$$\mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta) = \frac{1}{2}\mathrm{tr}\{\boldsymbol{Y}(\Psi^{-1} \otimes \Sigma^{-1})\boldsymbol{Y}'\} + h(\boldsymbol{\mu}) + f_0(\Theta) + \lambda \sum_{j=1}^{k} \omega_j \|\boldsymbol{\theta}_j\|.$$

Thus, we have

$$\frac{\partial}{\partial \boldsymbol{\theta}_r} \mathrm{RSS}_\lambda(\boldsymbol{\mu}, \Theta)\Big|_{\boldsymbol{\theta}_r = \hat{\boldsymbol{\theta}}_r} = (\boldsymbol{C}'\Psi^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})\hat{\boldsymbol{\theta}}_r + \frac{\lambda\omega_r \hat{\boldsymbol{\theta}}_r}{\|\hat{\boldsymbol{\theta}}_r\|} - \boldsymbol{g}_r(\mathbf{0}_{lq}),$$

where $\boldsymbol{g}_r(\mathbf{0}_{lq})$ is given by (26.8). The solution $\hat{\boldsymbol{\theta}}_r$ is the vector that makes the previous partial derivative equal to $\mathbf{0}_{lq}$, which, for given $\hat{\Psi}$, $\hat{\Sigma}$, and $\hat{\boldsymbol{\theta}}_j$ $(j \neq r)$, we obtain as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_r = &\left\{ \left(\boldsymbol{C}'\hat{\Psi}^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{X} \right) + \frac{\lambda\omega_r}{\|\hat{\boldsymbol{\theta}}_r\|} \boldsymbol{I}_{lq} \right\}^{-1} \\ &\times \left(\boldsymbol{C}'\hat{\Psi}^{-1} \otimes \boldsymbol{X}'\hat{\Sigma}^{-1} \right) \left\{ \boldsymbol{z}_r - (\boldsymbol{C} \otimes \boldsymbol{X})\hat{\Theta}'_{[-r]}\boldsymbol{v}_r \right\}. \end{aligned} \tag{26.11}$$

As will be explained in Sect. 26.2.4, these parameter matrices $\hat{\Psi}$, $\hat{\Sigma}$, and $\hat{\Theta}$ will be obtained, iteratively.

26.2.3 Optimizing λ

In order to determine possible candidate values for λ, we consider the maximum value of λ. From (26.10), $\hat{\boldsymbol{\theta}}_r = \mathbf{0}_{lq}$ $(r = 1, \ldots, k)$ when λ is sufficiently large. Thus, we consider the situation that all $\hat{\boldsymbol{\theta}}_r$ are $\mathbf{0}_{lq}$. When $\Theta = \mathbf{0}_k\mathbf{0}'_{lq}$, we obtain the estimators for Ψ and Σ as $\hat{\Psi}_0$ and $\hat{\Sigma}_0$, respectively. Then, the maximum value of λ is obtained as

$$\lambda_{\max} = \max_{j=1,\ldots,k} \|(\boldsymbol{C}'\hat{\Psi}_0 \otimes \boldsymbol{X}'\hat{\Sigma}_0^{-1})\boldsymbol{z}_j\|/\omega_j.$$

We use $\lambda_i = (3/4)^{100-i}\lambda_{\max}$ $(i = 1, \ldots, 100)$ to generate candidate values for λ, and obtain an estimate of Θ for each λ_i. Finally, we select the best λ_i by using an information criterion or threshold. The current study used BIC as the criterion as

$$\mathrm{BIC} = 2\ell_0(\hat{\boldsymbol{\mu}}, \hat{\Theta}, \hat{\Psi}, \hat{\Sigma}) + \{\#(\mathcal{J}) + 1\}\, lq \log n,$$

where ℓ_0 is the negative log-likelihood function given by (26.2), $\mathcal{J}$ is the active set defined by $\mathcal{J} = \{j \in \{1, \ldots, k\} | \, \hat{\boldsymbol{\theta}}_j \neq \mathbf{0}_{lq}\}$, and $\#(\mathcal{B})$ denotes the number of elements of the set $\mathcal{B}$.

26.2.4 Estimation Algorithm

We now have the necessary ingredients to compute estimates for all the unknown parameter matrices in the three-mode GMANOVA model (26.1), namely $\boldsymbol{\mu}$, Θ, Ψ, and Σ. The estimation algorithm proceeds iteratively as summarized below where variable s is the iteration index.

Initialization $(s = 0)$

Choose predefined convergence levels ϵ_1 and ϵ_2 as sufficiently small values (e.g., 10^{-4}). Choose the maximum number of iterations that one is willing to run (e.g., 10^3). Choose initial matrices for Ψ and Σ, denoted as $\hat{\Psi}^{(0)}$ and $\hat{\Sigma}^{(0)}$. Here, $\hat{\Psi}^{(0)}$ and $\hat{\Sigma}^{(0)}$ need to be positive definite matrices and the (1, 1)th element of $\hat{\Psi}^{(0)}$ is constrained to be 1 (this is necessary in order to avoid indeterminacy when computing $\Psi \otimes \Sigma$). Then we set $s = 1$. Using these initial matrices, we calculate $\hat{\boldsymbol{\mu}}^{(s-1)}$ and $\hat{\Theta}^{(s-1)}$

Iteration s $(s \geq 1)$

Step A Compute $\hat{\Theta}^{(s-1)}$ from (26.10) and (26.11) as follows, based on the estimates for Ψ and Σ given by $\hat{\Psi}^{(s-1)}$ and $\hat{\Sigma}^{(s-1)}$, respectively.

Step A-I Initialize $\hat{\Theta}^{(s-1)}$ as $\tilde{\Theta} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_k)'$, with $\hat{\Psi}^{(s-1)}$ as $\tilde{\Psi}$ and $\hat{\Sigma}^{(s-1)}$ as $\tilde{\Sigma}$. Here, $\tilde{\Theta}_{[-r]}$ is the $k \times lq$ matrix defined by replacing $\boldsymbol{\theta}_j$ in $\Theta_{[-r]}$ with $\tilde{\boldsymbol{\theta}}_j$ $(j \neq r)$.

Step A-II For each $r = 1, \ldots, k$, we first confirm whether $\boldsymbol{\theta}_r$ satisfies the following inequality:

$$\|\boldsymbol{w}_r\| \leq \lambda\omega_r,$$

where $\boldsymbol{w}_r = (\boldsymbol{C}'\tilde{\Psi}^{-1} \otimes \boldsymbol{X}'\tilde{\Sigma}^{-1})\{\boldsymbol{z}_r - (\boldsymbol{C} \otimes \boldsymbol{X})\tilde{\Theta}'_{[-r]}\boldsymbol{v}_r\}$. This inequality is derived from (26.10). If the inequality is satisfied, then update $\tilde{\boldsymbol{\theta}}_r = \mathbf{0}_{lq}$. Otherwise, proceed to Step A-II-i.

Step A-II Set $\boldsymbol{t}_0 = \tilde{\boldsymbol{\theta}}_r$ and $d = 1$.

Step A-II Using $\boldsymbol{t}_{d-1}$, calculate $\boldsymbol{t}_d$ with the following:

$$\boldsymbol{t}_d = \left(\boldsymbol{C}'\tilde{\Psi}^{-1}\boldsymbol{C} \otimes \boldsymbol{X}'\tilde{\Sigma}^{-1}\boldsymbol{X} + \frac{\lambda\omega_r}{\|\boldsymbol{t}_{d-1}\|}\boldsymbol{I}_{lq}\right)^{-1}\boldsymbol{w}_r,$$

which is derived from (26.11).

Step A-II If $|\|\boldsymbol{t}_d\|^2 - \|\boldsymbol{t}_{d-1}\|^2| < \epsilon_1 \|\boldsymbol{t}_{d-1}\|^2$, then update $\tilde{\boldsymbol{\theta}}_r$ as using $\boldsymbol{t}_d$. Otherwise, go to Step A-II-ii and set d as $(d+1)$.

Step A-III $\hat{\Theta}^{(s)}$ is derived as the updated $\tilde{\Theta}$.

Step B Compute $\hat{\Sigma}^{(s)}$ from $\hat{\Sigma}$ in (26.4) replacing $\boldsymbol{\mu}$, Θ, and Ψ with $\hat{\boldsymbol{\mu}}^{(s-1)}$, $\hat{\Theta}^{(s)}$, and $\hat{\Psi}^{(s-1)}$, respectively.

Step C Compute $\hat{\Psi}^{(s)}$ from $\hat{\Psi}$ in (26.4) replacing $\boldsymbol{\mu}$, Θ, and Σ with $\hat{\boldsymbol{\mu}}^{(s-1)}$, $\hat{\Theta}^{(s)}$, and $\hat{\Sigma}^{(s)}$, respectively.

Step D Compute $\hat{\boldsymbol{\mu}}^{(s)}$ from $\hat{\boldsymbol{\mu}}$ in (26.7) replacing Ψ and Σ with $\hat{\Psi}^{(s)}$ and $\hat{\Sigma}^{(s)}$, respectively.

Step E Divide $\hat{\Psi}^{(s)}$ by the $(1,1)$th component of $\hat{\Psi}^{(s)}$ and multiply $\hat{\Sigma}^{(s)}$ by the $(1,1)$th component of $\hat{\Psi}^{(s)}$.

Step F If $s > 1$ and $|\mathcal{L}_\lambda(s) - \mathcal{L}_\lambda(s-1)| < \epsilon_2 |\mathcal{L}_\lambda(s-1)|$, where $\mathcal{L}_\lambda(s) = \ell_\lambda(\hat{\boldsymbol{\mu}}^{(s)}, \hat{\Theta}^{(s)}, \hat{\Psi}^{(s)}, \hat{\Sigma}^{(s)})$, with ℓ_λ denoting the penalized negative log-likelihood function in (26.3), then stop the algorithm and retain $\hat{\boldsymbol{\mu}}^{(s)}$, $\hat{\Theta}^{(s)}$, $\hat{\Psi}^{(s)}$, and $\hat{\Sigma}^{(s)}$ as the estimates for $\boldsymbol{\mu}$, Θ, Ψ, and Σ, respectively. Otherwise, set s as $(s+1)$ and return to Step A.

The algorithm proceeds until the predefined convergence levels ϵ_1 and ϵ_2 are both reached or the maximum number of iterations is reached, whichever occurs first. Step A-II in this algorithm runs in ranking order of r. However, we may follow the steps that update $\boldsymbol{w}_r$ in an arbitrary order of r.

26.3 Numerical Experiments

Numerical experiments were conducted to examine the performance of our proposed method and those of existing variable selection approaches for comparison. The simulation data $\boldsymbol{Y}$ were generated from (26.1) using $\boldsymbol{\mu}$, $\boldsymbol{A}$, Θ, $\boldsymbol{C}$, $\boldsymbol{X}$, Ψ, and Σ produced by the following process:

- $\boldsymbol{\mu} = \mathbf{0}_{ql}$.
- The $\boldsymbol{A}$ was generated according to $\boldsymbol{A} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{A}_0 \Phi_k(\rho)^{1/2}$, where $\boldsymbol{J}_n = \mathbf{1}_n \mathbf{1}_n'/n$, $\Phi_k(\rho)$ is the $k \times k$ autoregressive correlation matrix with the (a,b)th element set as $\rho^{(a-b)}$ and $\boldsymbol{A}_0$ is an $n \times k$ matrix whose elements were independently generated from the uniform distribution over $(-1,1)$.
- $\Theta = \tau \Theta_0$ for $k \times lq$ matrix $\Theta_0 = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k^*}, \mathbf{0}_{ql}, \ldots, \mathbf{0}_{ql})'$ and τ as a fixed value. The elements of $\boldsymbol{\theta}_j$ for $j = 1, \ldots, k^*$ were generated independently from the uniform distribution over $(-1,1)$.
- The elements of $\boldsymbol{C}$ were generated independently from the uniform distribution over $(-1,1)$.
- The (a,b)th element of $\boldsymbol{X}$ was $a^{(b-1)}$ for $a = 1, \ldots, p$ and $b = 1, \ldots, q$.

- Positive definite Σ and Ψ which are for the covariance matrix of $\boldsymbol{Y}$ were generated using the `datasets.make_spd_matrix` function [4] in the scikit-learn library of Python package. The algorithm is as follows:

(a) Generate a $p \times p$ matrix $\boldsymbol{B}$ of independent entries following the uniform distribution over (0, 1).
(b) Define $p \times p$ orthogonal matrix $\boldsymbol{H}$ that diagonalizes $\boldsymbol{B}'\boldsymbol{B}$.
(c) Define $p \times p$ diagonal matrix $\boldsymbol{L}$ whose diagonal elements are independently generated from the uniform distribution over (0, 1) and construct $\boldsymbol{U} = \mathbf{1}_p \mathbf{1}_p' + \boldsymbol{L}$.
(d) Define $\boldsymbol{\mathcal{S}} = \boldsymbol{H}\boldsymbol{U}\boldsymbol{H}'$ and use $\boldsymbol{\mathcal{S}}$ as the positive definite matrix.

All the numerical experiments were done in Python, version 3.12. In the experiments, the values were fixed as follows: $(n, p, q, m, l, k, \rho, \tau) = (500, 5, 3, 3, 1, 50, 0.8, 0.15)$. However, k^* was manipulated to be 10, 20, 30, and 40, respectively.

To the generated data, we applied our proposed algorithm (G-Lasso), as well as existing variable selection approaches, including forward selection (Forward), backward elimination (Backward), forward-backward stepwise selection (F-B), and Kick-One-Out method (KOO; proposed by [9], named by [1]). BIC was used as the variable selection criterion for all methods. Four measures were used to evaluate performance: (1) Accuracy, the proportion of all correct identifications; (2) Sensitivity, the proportion of true positive identifications among all truly positive cases; (3) Specificity, the proportion of true negative identifications among all truly negative cases; and (4) Computational time in seconds. Each experiment was repeated 1,000 times for each tested approach and averages of the four measures were calculated. The results are summarized in Table 26.1, where all but computational time are given as percentages.

Looking at the results, we see that the KOO sometimes failed to select variables correctly, as reflected by Accuracy. The Forward, Backward, and F-B selection methods successfully selected variables, but with long computational times. On the other hand, our proposed method demonstrated that the Accuracy was either best or within 4% difference to the best accuracy method, while computational time were either the best or the second best.

26.4 Conclusion

In this study, we proposed the CDA of the group Lasso for selecting between individual explanatory variables in the three-mode GMANOVA model. Our numerical experiment showed that variable selection did not work perfectly in all tested cases. However, our proposed method demonstrated a good balance with respect to computational time and performance. Moreover, one advantage of implementing the group Lasso for the three-mode GMANOVA model is that if we want to select columns for explanatory variables to differentiate items among $\boldsymbol{C}$, or to select the relevant longitudinal time trend among $\boldsymbol{X}$, our proposed method can be extended easily,

Table 26.1 Results of numerical experiments

k^*	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Time (sec.)
10	G-Lasso	92.80	92.91	92.77	3.11
	KOO	89.90	49.70	99.95	3.05
	Forward	96.78	87.48	99.10	21.12
	Backward	96.89	87.52	99.24	65.28
	F-B	98.02	90.18	99.98	21.64
20	G-Lasso	90.24	97.04	85.71	3.09
	KOO	86.67	66.72	99.97	3.11
	Forward	91.15	86.46	94.28	36.47
	Backward	94.14	86.38	99.31	60.97
	F-B	94.17	86.06	99.58	38.74
30	G-Lasso	90.29	96.49	80.99	3.09
	KOO	73.33	55.58	99.95	2.87
	Forward	83.10	73.48	97.54	41.67
	Backward	83.32	72.87	99.01	53.77
	F-B	83.46	72.43	100.00	43.64
40	G-Lasso	91.14	96.33	70.37	3.12
	KOO	59.38	49.25	99.93	2.86
	Forward	72.94	67.94	92.94	48.21
	Backward	73.43	67.05	98.95	48.17
	F-B	72.04	65.06	100.00	52.06

compared to the other approaches. From this point of view, applying group Lasso to the three-mode GMANOVA model was found to be fruitful.

Acknowledgements The authors wish to thank two reviewers for their helpful comments. The third and last authors' research was partially supported by JSPS Bilateral Program Grant Number JPJSBP 120219927, and the last author's research was also partially supported by JSPS KAKENHI Grant Number 20H04151 and JSPS KAKENHI Grant Number 23H00809, and the first, third, and last authors' research was partially supported as the Institute of Statistical Mathematics Specially Promoted Research (2023-ISMCRP-4106).

References

1. Bai, Z., Fujikoshi, Y., Hu, J.: Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. (2018). arXiv:1810.12609
2. Harville, D.A.: Matrix Algebra from a Statistician's Perspective. Springer, New York (1997)
3. Horikawa, K., Nagai, I., Monden, R., Yanagihara, H.: Estimation algorithms for MLE of three-mode GMANOVA model with Kronecker product covariance matrix. Smart Innov. Syst. Tec. **352**, 203–213 (2023). https://doi.org/10.1007/978-981-99-2969-6_18

4. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). https://www.mlr.org/papers/v12/pedregosa11a.html
5. Potthoff, R.F., Roy, S.N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**, 313–326 (1964). https://doi.org/10.2307/2334137
6. Wang, H., Leng, C.: A note on adaptive group lasso. Comput. Stat. Data Anal. **52**, 5277–5286 (2008). https://doi.org/10.1016/j.csda.2008.05.006
7. Yanagihara, H., Oda, R.: Coordinate descent algorithm for normal-likelihood-based group Lasso in multivariate linear regression. Smart Innov. Syst. Tec. **238**, 429–439 (2021). https://doi.org/10.1007/978-981-16-2765-1_36
8. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. Ser. B **68**, 49–67 (2006). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
9. Zhao, L.C., Krishnaiah, P.R., Bai, Z.D.: On detection of the number of signals in presence of white noise. J. Multivariate Anal. **20**, 1–25 (1986). https://doi.org/10.1111/j.1467-9868.2005.00532.x
10. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 1418–1429 (2006). https://doi.org/10.1198/016214506000000735

Chapter 27
Poisson Regression with Categorical Explanatory Variables via Lasso Using the Median as a Baseline

Mariko Yamamura, Mineaki Ohishi, and Hirokazu Yanagihara

Abstract This study shows that the objective function of a Lasso Poisson regression with categorical explanatory variables can be explicitly minimized in one direction. The coordinate descent method is used to obtain the optimal coefficients. Because estimates of the coefficients for each category of a categorical variable and the interpretation of these estimates depend on the baseline, it is crucial to identify the baseline category, which in the Lasso regression is the one whose coefficient is estimated to be zero. Theoretical clarifications and numerical experiments show that the baseline corresponds to a category with a coefficient equal to the weighted median of the coefficients of the categorical variable.

27.1 Introduction

Unlike in linear regression, in Poisson regression, the objective function cannot be explicitly minimized with respect to parameters. Additionally, unlike in logistic regression, it is not possible to derive an update equation by linear approximation that can be solved explicitly, always reducing the value of the objective function. In other words, the analysis requires searching for solutions through numerical comparisons. However, [7] demonstrated in fused Lasso (see [4]) Poisson regression that when all the explanatory variables are categorical, the objective function can be explicitly minimized with respect to parameters in one direction. They proposed an estimation approach using the coordinate descent method (CDM, see [1]), which leads to an optimal solution by sequentially minimizing the objective function in one direction. We present a method for explicitly determining the parameters of a Lasso (see [3]) Poisson regression when all the explanatory variables are categorical.

M. Yamamura · H. Yanagihara (✉)
Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8526, Japan
e-mail: yanagi-hiro@hiroshima-u.ac.jp

M. Ohishi
Center for Data-driven Science and Artificial Intelligence, Tohoku University, Sendai, Miyagi 980-8576, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_27

The Lasso method is attractive because it allows for the estimation of coefficients to be completely zero. This means that explanatory variables with coefficients estimated to be zero do not need to be included in the model, making Lasso a useful tool for variable selection. Lasso has proven effective for variable selection in categorical variable cases, such as in spatial analysis in geographic information systems (GIS). As shown in [6], for a categorical variable where each region is a category, variable selection for the categories is performed in Lasso, while neighboring regions with the same effect are fused as the same category in fused Lasso.

Consider the implications of analyzing categorical variables in a Lasso regression. In statistical analyses without Lasso, such as linear regression, a category of the categorical variable is defined as the baseline and omitted from the model due to multicollinearity. Because the estimates of the coefficients corresponding to the categories of the categorical variable are based on the category used as the baseline, changing the baseline category will result in different estimates. Therefore, a data analyst usually determines the baseline category so that the interpretation of the analysis results is clear. In Lasso regression, estimating the coefficients of the categorical variable to be zero is equivalent to setting the baseline. The remaining question is how to determine the categories with zero coefficients, that is, how to establish the baseline in the Lasso regression. Understanding which category serves as the baseline is crucial for interpreting the analysis results. In this paper, we present a method for explicitly determining the solution and describe the categories for which the coefficients are estimated to be zero and thus become the baseline category.

The remainder of the paper proceeds are as follows: Sect. 27.2 describes the proposed method. Here, we show an objective function that explicitly determines the solution and derive a baseline category. The results of numerical experiments are provided in Sect. 27.3. Conclusions are presented in Sect. 27.4.

27.2 Model and Estimation

Assume the count data $y_i \in \{z \in \mathbb{Z} | \, z \geq 0\}, i = 1, \ldots, n$ or the rate $y_i/m_i, m_i > 0$ follows Poisson distribution with parameter $\theta_i, i = 1, \ldots, n$. By using a logarithmic link function, we further assume that $\theta_i = m_i \exp(\mu + \boldsymbol{x}_i'\boldsymbol{\beta})$, where μ is a constant term, $\boldsymbol{x}_i$ is a p-dimensional explanatory variable vector for sample i, $\boldsymbol{\beta}$ is the p-dimensional coefficient vector, and "$'$" denotes transposition. Vectors and matrices are shown in bold, and scalars are shown in normal typeface. The explanatory variable $\boldsymbol{x}_i$ consists of k kinds of categorical variables, denoted $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}', \ldots, \boldsymbol{x}_{i,k}')'$. The categorical variable $\boldsymbol{x}_{i,j} = (x_{i,j,1}, \ldots x_{i,j,p_j})', \ j = 1, \ldots, k$ contains p_j categories, $\sum_{j=1}^{k} p_j = p$. The coefficients corresponding to $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i,j}$ are denoted by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k')'$ and $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,p_j})'$, respectively.

The negative log-likelihood function for Poisson regression is

$$F(\mu, \boldsymbol{\beta}) = -\sum_{i=1}^{n} \left\{ y_i \log m_i + y_i \left(\mu + \boldsymbol{x}_i' \boldsymbol{\beta}\right) - m_i \exp\left(\mu + \boldsymbol{x}_i' \boldsymbol{\beta}\right) - \log y_i! \right\}$$
$$= F_j\left(\mu, \boldsymbol{\beta}_j\right) - \sum_{i=1}^{n} y_i \eta_{i(-j)} + \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} y_i \log m_i, \tag{27.1}$$

where $\eta_{i(-j)} = \sum_{\ell \neq j}^{k} \boldsymbol{x}_{i,\ell}' \boldsymbol{\beta}_\ell = \boldsymbol{x}_i' \boldsymbol{\beta} - \boldsymbol{x}_{i,j}' \boldsymbol{\beta}_j$. Focusing on categorical variable j and letting $\alpha_{i,j} = \exp(\eta_{i(-j)})$, $F_j(\mu, \boldsymbol{\beta}_j)$ in (27.1) is a different form for $F(\mu, \boldsymbol{\beta})$ as

$$F_j(\mu, \boldsymbol{\beta}_j) = -\sum_{i=1}^{n} y_i \left(\mu + \boldsymbol{x}_{i,j}' \boldsymbol{\beta}_j\right) + \sum_{i=1}^{n} m_i \alpha_{i,j} \exp\left(\mu + \boldsymbol{x}_{i,j}' \boldsymbol{\beta}_j\right). \tag{27.2}$$

Consider representing the penalized negative log-likelihood function of $\boldsymbol{\beta}$ by a function of $\boldsymbol{\beta}_j$ and minimizing it to obtain the optimal solution. For the component $\beta_{j,\ell}$, $\ell = 1, \ldots, p_j$, of $\boldsymbol{\beta}_j$, we present an explicitly minimizable objective function and use the CDM to find the optimal solution for $\boldsymbol{\beta}_j$ and hence $\boldsymbol{\beta}$.

To find the solution to $\beta_{j,\ell}$, (27.2) is transformed into an expression in terms of categories 1 through p_j. As preparation, let $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\alpha}_j = (\alpha_{1,j}, \ldots, \alpha_{n,j})'$, $\boldsymbol{X}_j = (\boldsymbol{x}_{1,j}, \ldots, \boldsymbol{x}_{n,j})'$, and $\bar{\boldsymbol{y}}_j = (\bar{y}_{j,1}, \ldots, \bar{y}_{j,p_j})' = n^{-1} \boldsymbol{X}_j' \boldsymbol{y}$. Note that $\boldsymbol{X}_j$ is an $n \times p_j$ matrix and $\bar{\boldsymbol{y}}_j$ is a p_j-dimensional vector. Here, since $\boldsymbol{X}_j \mathbf{1}_{p_j} = \mathbf{1}_n$, the mean of y_i can be expressed as $\bar{y} = \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell}$, where $\mathbf{1}_n$ is an n-dimensional vector with component 1. Next, rewrite (27.2) as

$$F_j(\mu, \boldsymbol{\beta}_j) = -n\bar{y}\mu - n \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell} \beta_{j,\ell} + n e^{\mu} \sum_{\ell=1}^{p_j} a_{j,\ell} \exp(\beta_{j,\ell}), \tag{27.3}$$

where $a_{j,\ell}$ is a component of $\boldsymbol{a}_j = (a_{j,1}, \ldots, a_{j,p_j})' = n^{-1} \boldsymbol{X}_j' \boldsymbol{D}_m \boldsymbol{\alpha}_j$, and $\boldsymbol{D}_m = \mathrm{diag}(m_1, \ldots, m_n)$.

27.2.1 Objective Function

Let $\lambda > 0$ be the regularization parameter and $w_{j,\ell}$ be the adaptive Lasso weight in [9]. Then, the penalized negative log-likelihood function of $\boldsymbol{\beta}$, i.e., the objective function for the optimization of μ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$, is

$$f(\mu, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k | \lambda) = F(\mu, \boldsymbol{\beta}) + n\lambda \sum_{j=1}^{k} \sum_{\ell=1}^{p_j} w_{j,\ell} |\beta_{j,\ell}|, \tag{27.4}$$

where $w_{j,\ell} = 1/|\tilde{\beta}_{j,\ell}|$, and $\tilde{\beta}_{j,\ell}$ is the maximum likelihood estimator (MLE) of $\beta_{j,\ell}$ under the constraints $\sum_{\ell=1}^{p_j} \beta_{j,\ell} = 0$. Since there is no penalty for μ in (27.4), μ is obtained by differentiating (27.1) by μ with $\boldsymbol{\beta}$ given. That is

$$\mu = \log\left\{ \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \right\}. \tag{27.5}$$

From (27.3), the objective function for the optimization of $\boldsymbol{\beta}_j$ is

$$f_j(\mu, \boldsymbol{\beta}_j|\lambda) = -\mu\bar{y} - \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell}\beta_{j,\ell} + e^{\mu} \sum_{\ell=1}^{p_j} a_{j,\ell} \exp(\beta_{j,\ell}) + \lambda \sum_{\ell=1}^{p_j} w_{j,\ell}|\beta_{j,\ell}|. \tag{27.6}$$

Since excluding n from (27.3) has no effect on the optimization, n is omitted in (27.6).

For estimation using the CDM, the optimal solution is obtained by solving (27.6) for $\beta_{j,1}, \ldots, \beta_{j,p_j}$ in one direction. The derivative of $f_j(\mu, \boldsymbol{\beta}_j|\lambda)$ at points other than $\beta_{j,\ell} = 0$ is

$$\dot{f}_j(\mu, \boldsymbol{\beta}_j|\lambda) = \frac{\partial}{\partial \beta_{j,\ell}} f_j(\mu, \boldsymbol{\beta}_j|\lambda) = -\bar{y}_{j,\ell} + e^{\mu} a_{j,\ell} \exp(\beta_{j,\ell}) + \lambda w_{j,\ell} \mathrm{sign}(\beta_{j,\ell}). \tag{27.7}$$

The condition under which (27.7) is minimized at $\beta_{j,\ell} = 0$ is as follows:

$$\lim_{\beta_{j,\ell} \to 0+} \dot{f}_j(\mu, \boldsymbol{\beta}_j|\lambda) = -\bar{y}_{j,\ell} + e^{\mu} a_{j,\ell} + \lambda w_{j,\ell} \geq 0, \tag{27.8}$$

$$\lim_{\beta_{j,\ell} \to 0-} \dot{f}_j(\mu, \boldsymbol{\beta}_j|\lambda) = -\bar{y}_{j,\ell} + e^{\mu} a_{j,\ell} - \lambda w_{j,\ell} \leq 0, \tag{27.9}$$

Expressions (27.8) and (27.9) can be written collectively as

$$|\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell}| \leq \lambda w_{j,\ell} \Longleftrightarrow \text{(27.6) is minimum at } \beta_{j,\ell} = 0. \tag{27.10}$$

Consider the minimum solution when (27.10) does not hold. By putting (27.7) equal to 0 when $\beta_{j,\ell} \neq 0$, (27.7) can be written as

$$\begin{aligned} e^{\mu} a_{j,\ell}\{\exp(\beta_{j,\ell}) - 1\} &= \bar{y}_{j,\ell} - e^{\mu} a_{j,\ell} - \mathrm{sign}(\beta_{j,\ell})\lambda w_{j,\ell} \\ &= \mathrm{sign}(\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell})\{|\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell}| - \lambda w_{j,\ell}\}. \end{aligned} \tag{27.11}$$

Since (27.10) does not now hold, $|\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell}| > \lambda w_{j,\ell}$, where $\lambda w_{j,\ell} \geq 0$; therefore,

$$\begin{aligned} &\mathrm{sign}\ \{\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell} - \mathrm{sign}(\beta_{j,\ell})\lambda w_{j,\ell}\} = \mathrm{sign}(\beta_{j,\ell}) \\ \Longleftrightarrow\ &\mathrm{sign}\ (\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell}) = \mathrm{sign}(\beta_{j,\ell}). \end{aligned}$$

Solving (27.11) for $\beta_{j,\ell}$ and using the solution and the condition in (27.10), we obtain

$$\beta_{j,\ell} = I(|\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell}| > \lambda w_{j,\ell}) \log\left\{ \frac{\bar{y}_{j,\ell} - \text{sign}(\bar{y}_{j,\ell} - e^{\mu} a_{j,\ell})\lambda w_{j,\ell}}{e^{\mu} a_{j,\ell}} \right\}, \quad (27.12)$$

where $I(A)$ is the indicator function, that is, $I(A) = 1$ if A is true and $I(A) = 0$ if A is false.

27.2.2 Baseline in the Categorical Variable

We can theorize which of the $\beta_{j,1}, \ldots, \beta_{j,p_j}$ are estimated to be zero by Lasso, i.e., which category would be considered the baseline. In the Lasso regression model, as in the model without Lasso, the analyst is able to exclude from the explanatory variables those categories that he or she identifies as the baseline. In this case, however, all categories are used, including the baseline category.

Given $z_1, \ldots, z_q \in \mathbb{R}$, it is well known that the d that minimizes the sum of squared deviation, $\sum_{i=1}^{q} |z_i - d|$, is the median of $z_1, \ldots, z_q$. We apply this to the penalty term in (27.6) and consider centering $\boldsymbol{\beta}_j$ at a certain value of δ_j. Given $\mu = \hat{\mu}$, let $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,p_j})'$ be the $\boldsymbol{\beta}_j$ that minimizes (27.6), and thus $f_j(\hat{\mu}, \boldsymbol{\beta}_j|\lambda) \geq f_j(\hat{\mu}, \hat{\boldsymbol{\beta}}_j|\lambda)$. If we let

$$\hat{\delta}_j = \arg\min_{\delta_j} \sum_{\ell=1}^{p_j} w_{j,\ell} |\hat{\beta}_{j,\ell} - \delta_j|, \quad (27.13)$$

then $\sum_{\ell=1}^{p_j} w_{j,\ell} |\hat{\beta}_{j,\ell}| \geq \sum_{\ell=1}^{p_j} w_{j,\ell} |\hat{\beta}_{j,\ell} - \hat{\delta}_j|$. Note $\bar{y} = \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell}$, so we obtain,

$$\begin{aligned} f_j(\hat{\mu}, \hat{\boldsymbol{\beta}}_j|\lambda) &\geq -\hat{\mu}\bar{y} - \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell} \hat{\beta}_{j,\ell} + e^{\hat{\mu}} \sum_{\ell=1}^{p_j} a_{j,\ell} \exp(\hat{\beta}_{j,\ell}) + \lambda \sum_{\ell=1}^{p_j} w_{j,\ell} |\hat{\beta}_{j,\ell} - \hat{\delta}_j| \\ &= -(\hat{\mu} + \hat{\delta}_j)\bar{y} - \sum_{\ell=1}^{p_j} \bar{y}_{j,\ell} (\hat{\beta}_{j,\ell} - \hat{\delta}_j) + e^{\hat{\mu} + \hat{\delta}_j} \sum_{\ell=1}^{p_j} a_{j,\ell} \exp(\hat{\beta}_{j,\ell} - \hat{\delta}_j) \\ &\quad + \lambda \sum_{\ell=1}^{p_j} w_{j,\ell} |\hat{\beta}_{j,\ell} - \hat{\delta}_j| = f_j(\hat{\mu} + \hat{\delta}_j, \hat{\boldsymbol{\beta}}_j - \hat{\delta}_j \mathbf{1}_{p_j}|\lambda). \end{aligned} \quad (27.14)$$

Let $(\hat{\mu}, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k)$ be the minimizer of (27.4) obtained by the CDM using (27.5) and (27.12). According to (27.14), it can be concluded that

$$f(\hat{\mu}, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k|\lambda) \geq f(\hat{\mu} + (\hat{\delta}_1 + \cdots + \hat{\delta}_k), \hat{\boldsymbol{\beta}}_1 - \hat{\delta}_1 \mathbf{1}_{p_1}, \ldots, \hat{\boldsymbol{\beta}}_k - \hat{\delta}_k \mathbf{1}_{p_k}|\lambda). \quad (27.15)$$

[5] reported that the CDM can minimize non-differentiable functions. This indicates that $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$, which minimize (27.4), should be centered at $\hat{\delta}_1, \ldots, \hat{\delta}_k$, respectively. If the estimate is zero, $\hat{\beta}_{j,\ell} = \hat{\delta}_j$, meaning that the baseline is a category with a coefficient value equal to $\hat{\delta}_j$.

The $\hat{\delta}_j$ value can be obtained as follows: Let $\hat{\beta}_{j,(1)}, \ldots, \hat{\beta}_{j,(p_j)}$ be the new sequence created by rearranging $\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,p_j}$ in ascending order, let $w_{j,(\ell)}$ be the adaptive Lasso weight corresponding to $\hat{\beta}_{j,(\ell)}$. If we then let $\delta_j \in (\beta_{j,(q)}, \beta_{j,(q+1)}]$, $q = 1, \ldots, p_j - 1$, the summation in (27.13) can be rewritten as

$$\begin{aligned}\sum_{\ell=1}^{p_j} w_{j,\ell}|\hat{\beta}_{j,\ell} - \delta_j| &= \sum_{\ell=1}^{p_j} w_{j,(\ell)}|\hat{\beta}_{j,(\ell)} - \delta_j| \\ &= \left(\sum_{\ell=1}^{q} w_{j,(\ell)} - \sum_{\ell=q+1}^{p_j} w_{j,(\ell)}\right)\delta_j - \sum_{\ell=1}^{q} w_{j,(\ell)}\hat{\beta}_{j,(\ell)} + \sum_{\ell=q+1}^{p_j} w_{j,(\ell)}\hat{\beta}_{j,(\ell)}. \quad (27.16)\end{aligned}$$

Looking at the terms in parentheses in (27.16), let

$$\Omega_{j,0} = -\Omega_{j,p_j}, \quad \Omega_{j,p_j} = \sum_{\ell=1}^{p_j} w_{j,(\ell)}, \quad \Omega_{j,q} = \sum_{\ell=1}^{q} w_{j,(\ell)} - \sum_{\ell=q+1}^{p_j} w_{j,(\ell)}. \quad (27.17)$$

From (27.17), we have

$$\begin{aligned}&\exists q \in \{0, 1, \ldots, p_j\} \text{ s.t. } \Omega_{j,q} < 0,\ \Omega_{j,q+1} > 0 \ \Rightarrow\ \hat{\delta}_j = \hat{\beta}_{j,(q+1)}, \\ &\exists q \in \{0, 1, \ldots, p_j\} \text{ s.t. } \Omega_{j,q} = 0 \ \Rightarrow\ \hat{\delta}_j \in \left[\hat{\beta}_{j,(q)}, \hat{\beta}_{j,(q+1)}\right]. \quad (27.18)\end{aligned}$$

Since $w_{j,\ell} > 0$, there always exists q. From (27.18), $\hat{\delta}_j$ is the "weighted median" of $\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,p_j}$ with adaptive Lasso weights. If $w_{j,1} = \cdots = w_{j,p_j}$, $\hat{\delta}_j$ is the median of $\hat{\beta}_{j,1}, \ldots \hat{\beta}_{j,p_j}$.

27.3 Simulation Study

Numerical experiments were performed for the proposed method. Here, we call the optimization method that minimizes (27.6) the "Non-Centered Optimization Method (NCOM)", while the optimization method that minimizes (27.14), which centers $\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,p_j}$ at $\hat{\delta}_j$, is called the "Centered Optimization Method (COM)". The estimation algorithms are shown in Tables 27.1 and 27.2.

The numerical experiments answer the following three questions: (i) Does (27.15) actually hold?, (ii) What is the difference in computation time between NCOM and

Table 27.1 Algorithm for NCOM

Input: $\boldsymbol{\beta}_{\text{int}}$ (Initial values)
Output: $\boldsymbol{\beta}_{\text{opt}}$, μ_{opt} (Optimal values)
1: Compute μ from (27.5) using $\boldsymbol{\beta}_{\text{int}}$
2: **for** $j = 1, \ldots, k$
3: **for** $\ell = 1, \ldots, p_j$
4: Update $\beta_{j,\ell}$ from (27.12)
5: **end for**
6: **end for**
7: Update μ from (27.5) using updated $\boldsymbol{\beta}$
Repeat 2-7 *until* μ *and* $\boldsymbol{\beta}$ *are converged.*
The converged results are μ_{opt} *and* $\boldsymbol{\beta}_{\text{opt}}$.

Table 27.2 Algorithm for COM

Input: $\boldsymbol{\beta}_{\text{int}}$ (Initial values)
Output: $\boldsymbol{\beta}_{\text{opt}}$, μ_{opt} (Optimal values)
1: Compute μ from (27.5) using $\boldsymbol{\beta}_{\text{int}}$.
2: **for** $j = 1, \ldots, k$
3: **for** $\ell = 1, \ldots, p_j$
4: Update $\boldsymbol{\beta}_{j,\ell}$ from (27.12).
5: **end for**
6: Compute δ_j from (27.13)
7: Update $\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j - \delta_j \mathbf{1}_{p_j}$ from (27.14)
8: Update μ with $\mu + \delta_j$ from (27.14)
9: **end for**
10: Update μ from (27.5) using updated $\boldsymbol{\beta}$
Repeat 2-10 *until* μ *and* $\boldsymbol{\beta}$ *are converged.*
The converged results are μ_{opt} *and* $\boldsymbol{\beta}_{\text{opt}}$.

COM?, and (iii) Will a model that omits baseline categories determined by an analyst yield the same results as a model that does not omit them? The value of δ_j is not necessarily the same when the baseline category is omitted from the categorical variable as when it is not. Therefore, it is questionable whether omitting the baseline category will yield the same estimates as not omitting it.

There are five categorical variables, each consisting of five categories. The true values of the coefficients are $\boldsymbol{\beta}_1 = (-1, -0.5, 0, 1, 2)'$, $\boldsymbol{\beta}_2 = (-1, 0, 0, 1, 2)'$, $\boldsymbol{\beta}_3 = (-1, 0, 0, 0, 2)'$,and $\boldsymbol{\beta}_4 = \boldsymbol{\beta}_5 = (0, 0, 0, 0, 0)'$. Categorical variables are randomly generated based on a multinomial distribution with all cell probabilities equal to $1/5$. These variables are used to generate y_i, which follows a Poisson distribution, and to prepare data for $n = 100$, 300, and 500 samples. However, $m_i = 1$, $i = 1, \ldots, n$.

Since λ is an unknown value, we enumerate the candidates for λ, obtain an estimate of $\boldsymbol{\beta}$ under each candidate, and select the best estimate among them using BIC. If we know the minimum and maximum values that the candidate λ can take, we can focus only on the values in between. As shown in standard textbooks on sparse estimation, e.g., [2], it is known that $\boldsymbol{\beta}$ shrinks as the value of λ increases. The minimum value of λ is set when all parameters, μ and $\boldsymbol{\beta}$, are estimated and $\lambda_{\min} = 0$. On the other hand, the maximum value of λ is the value when there is only one parameter, μ. In this case, $\boldsymbol{\beta} = \mathbf{0}$. Therefore, based on (27.10), we have $\lambda_{\max} = \max_{(j,\ell)\in\mathcal{J}} |\bar{y}_{j,\ell} - \exp(\hat{\mu}_{\max}) a^{\star}_{j,\ell}|/w_{j,\ell}$, where $\mathcal{J} = \{(1,1), \ldots, (1,p_1), \ldots, (k,1), \ldots, (k,p_k)\}$, $\hat{\mu}_{\max} = \log(n\bar{y}/\sum_{i=1}^{n} m_i)$, and $\boldsymbol{a}^{\star}_j = n^{-1}\boldsymbol{X}'_j \boldsymbol{D}_m \mathbf{1}_n$. For the latter, $\boldsymbol{a}_j = n^{-1}\boldsymbol{X}'_j \boldsymbol{D}_m \boldsymbol{\alpha}_j$ when $\boldsymbol{\beta}_\ell = \mathbf{0}_{p_\ell}$, $\ell \neq j$.

Five estimation methods are provided for comparison. The (C) and (N) are optimized by COM and NCOM, respectively, without omitting the baseline category for each categorical variable. On the other hand, the (B1), (B2), and (B3) are optimized by NCOM, omitting the baseline category. In (B1), (B2), and (B3), the categories corresponding to the "minimum value, $\min(\tilde{\beta}_{j,1}, \ldots, \tilde{\beta}_{j,p_j})$", "maximum value, $\max(\tilde{\beta}_{j,1}, \ldots, \tilde{\beta}_{j,p_j})$", and "closest value to the mean, $\sum_{\ell=1}^{p_j} \tilde{\beta}_{j,\ell}/p_j$" in the MLE, respectively, were omitted as the baseline.

The estimation was repeated 1000 times for each method, from (C) to (B3). For the iterative calculations by the CDM, the convergence of the estimates is determined as $\max\{(\hat{\mu}^{\text{new}} - \hat{\mu}^{\text{old}})^2, \max_{(j,\ell)\in\mathcal{J}}(\hat{\beta}^{\text{new}}_{j,\ell} - \hat{\beta}^{\text{old}}_{j,\ell})^2\} \leq \varepsilon \in (10^{-5}, 10^{-10})$. The convergence is lenient (LC) when $\varepsilon = 10^{-5}$, and is strict (SC) when $\varepsilon = 10^{-10}$. The subscripts "new" and "old" are the parameter estimates obtained by iterative computation, where "new" is updated and taken as the estimated result if it is smaller than the ε, and "old" is the estimate before "new", i.e., before the update.

Figure 27.1 compares the results of estimation using data with a sample size of 100 between (C) and (N). The horizontal axis displays the coefficient $\beta_{j,\ell}$, ordered from $\beta_{1,1}$ to $\beta_{5,5}$, with values ranging from 1.1 to 5.1 indicating the subscript of $\beta_{j,\ell}$ in the corresponding scale. The vertical axis shows the absolute value of the difference between the estimates of $\beta_{j,\ell}$ obtained from (C) and (N); the boxplot displays the 1000 estimated results. Given the proximity of the values on the vertical axis to zero in SC, there is minimal distinction between (C) and (N) when the convergence condition is strict.

Figure 27.2 compares the minimum objective function values for (C) and (N) obtained from the analyses in Fig. 27.1. The horizontal axis displays the 100 regularization parameters λ in ascending order from left to right. The vertical axis shows the difference (N)−(C) in terms of the minimum objective function value. The boxplot displays the 1000 estimated results for each λ, with little difference observed between (C) and (N) for both LC and SC when λ takes small or large values. However, in the range of λ where the difference is large, many of the values on the vertical axis are positive, indicating that the minimum value of the objective function in (C) is smaller than that in (N). However, even when the differences appear significant, the values on the vertical axis are small, indicating minimal variation in SC.

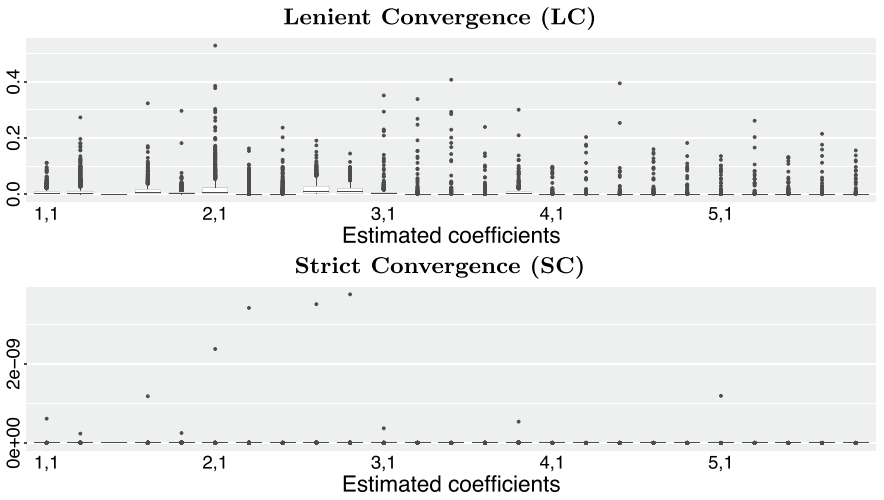


Fig. 27.1 Difference between the estimated values of the coefficients: |(N)−(C)|

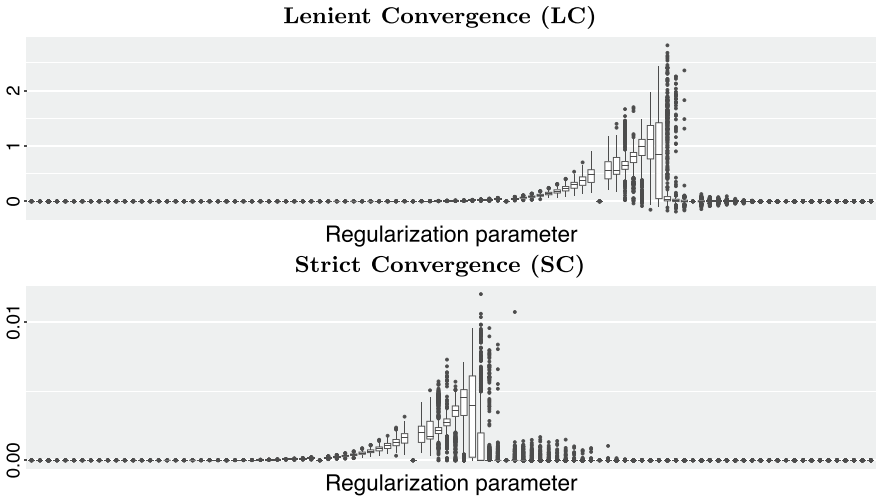


Fig. 27.2 Difference between the minimums of the objective functions: (N)−(C)

Table 27.3 in parentheses compares the computation times for LC and SC for (C) and (N). The computation times for (C) and (N) are almost identical for LC. However, (C) is significantly faster than (N) for SC, and this difference becomes more pronounced as the sample size increases. Table 27.3 not in parentheses compares the accuracy of variable selection. The value shown is the probability, in percentage terms, that the estimated result includes the true categorical variables, i.e., the probability that β_1, β_2, and β_3 are included in the model and β_4 and β_5 are not. The difference between LC and SC is small and increases as the sample size increases.

Table 27.3 Accuracy of variable selection (%) and computation time (sec.)

n	LC					SC				
	(C)	(N)	(B1)	(B2)	(B3)	(C)	(N)	(B1)	(B2)	(B3)
100	72.9 (0.12)	74.9 (0.07)	30.1	72.1	66.3	72.8 (1.01)	72.8 (8.04)	27.3	72.1	65.8
300	93.7 (0.14)	95.3 (0.11)	56.8	89.8	85.5	93.1 (0.92)	93.1 (15.65)	51.4	89.4	84.1
500	97.8 (0.19)	99.0 (0.19)	73.9	92.5	92.1	97.9 (1.07)	97.9 (25.45)	64.9	90.6	91.9

However, for the largest sample size of 500, (B1) is 73.9% for LC and 64.9% for SC, indicating that the probability of selecting a true categorical variable is not high. The probabilities of (B2) and (B3) are not as high as those of (C) and (N), but they are still significant, being over 90% in a sample of 500. However, as the sample size decreases, the probabilities of (B2) and (B3) decrease substantially compared to those of (C) and (N).

27.4 Conclusion

Figure 27.1 illustrates that the estimates of (C) and (N) are nearly identical in SC. Additionally, Fig. 27.2 shows that the difference between the minimum objective function values of (C) and (N) is smaller for SC than for LC. This could be due to the fact that the minimum objective function value of (N) is even smaller in SC than in LC, resulting in a value closer to that of (C). In other words, the minimization of NCOM has progressed to the point where the inequality in (27.15) can be expressed by the equal sign, which implies that the estimate of NCOM is equal to the estimate of COM. Therefore, for (i), (27.15) actually holds, and it is particularly important to strictly adhere to the convergence condition. Thus, the estimation results obtained from Lasso are centered on the weighted median. Consequently, the category with the weighted median as the coefficient is the baseline.

For (ii), Table 27.3 shows that the computation time for (C) is significantly shorter than that for (N). If the sample size is large, the estimation can be conducted using COM rather than NCOM. Regarding (iii), the results of (B1), (B2), and (B3) in Table 27.3 demonstrate that the probability of selecting the wrong model increases when the baseline category is omitted in the NCOM. Furthermore, this probability is particularly high with small sample sizes. Therefore, when using Lasso, it is recommended that all categories be included in the model for estimation and that the baselines not be excluded prior to estimation.

We describe a contribution unrelated to (i), (ii), and (iii). In Table 27.3, it is noteworthy that the categorical variables $\boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$ have coefficients that take the true value of 0 and are correctly omitted from the estimation results for (C) and (N)

with sample sizes of 300 and 500 and probabilities greater than 90%. The group Lasso proposed in [8] can be used to select or not select a categorical variable, but the solution is not explicitly determinable. The contribution of this study is that the selection is accomplished by explicitly minimizing the objective function.

Acknowledgements The authors wish to thank two reviewers for their helpful comments. This research was supported by JSPS Bilateral Program Grant Number JPJSBP 120219927 and JSPS KAKENHI Grant Number 20H04151. The second author's research was partially supported by JSPS KAKENHI Grant Number 21K13834 and the ISM Specially Promoted Research (2023-ISMCRP-4105). The third author's research was partially supported by JSPS KAKENHI Grant Number 23H00809.

References

1. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22 (2010). https://doi.org/10.18637/jss.v033.i01
2. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton, FL (2015)
3. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B. Methodol. **58**, 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
4. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused Lasso. J. R. Stat. Soc. Series B. Stat. Methodol. **67**, 91–108 (2005). https://doi.org/10.1111/j.1467-9868.2005.00490.x
5. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 475–494 (2001). https://doi.org/10.1023/A:1017501703105
6. Wang, H., Rodríguez, A.: Identifying pediatric cancer clusters in Florida using log-linear models and generalized Lasso penalties. Stat. Publ. Policy **1**, 86–96 (2014). https://doi.org/10.1080/2330443X.2014.960120
7. Yamamura, M., Ohishi, M., Yanagihara, H.: Additive Poisson regression via forced categorical covariates and generalized fused Lasso. Procedia Comput. Sci. **225**, 1987–1996 (2023). https://doi.org/10.1016/j.procs.2023.10.189
8. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B: Stat. Methodol. **68**:49–67 (2006). https://doi.org/10.1111/j.1467-9868.2005.00532.x
9. Zou, H.: The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006). https://doi.org/10.1198/016214506000000735

Chapter 28
Generalized Triply Robust Information Criterion

Yoshiyuki Ninomiya

Abstract In semiparametric propensity score analysis, which is a standard in causal inference, we consider incorporating a loss function that is robust to outliers. First, we confirm that the estimation employing covariate balancing is still doubly robust. An over-identified case, in which the number of moment conditions considered in covariate balancing is greater than the number of parameters, is also addressed. Then, we propose a generalized triply robust information criterion, gTRIC, which is valid in this setting. Numerical experiments show that gTRIC performs better than existing information criteria in the presence of outliers.

28.1 Introduction

Akaike information criterion (AIC; Akaike [2]) has become a basic tool for statistical analysis and has since been developed in various directions. On the other hand, AIC has not been fully developed in specific areas such as singular model analysis and change point analysis. Even in such a specific area, the development of AIC is highly meaningful in cases where the rigorous reevaluation of AIC can significantly change analysis results, as compared to the formal use of AIC. This paper deals with causal inference in which reevaluating AIC can significantly change the results. In Rubin's causal model, the outcome variables are considered to potentially exist as many as the number of treatments, but only the outcome variable corresponding to the assigned treatment is observed. If the outcome and assignment variables are confounded and estimation is done naively with the marginal likelihood without taking the confounding into account, the estimate will have a bias. The bias can be removed if the confounding can be modeled correctly, but instead of such difficult modeling, a semiparametric approach such as inverse-probability-weighted estimation (Robins et al. [10]) or doubly robust estimation (Scharfstein et al. [12]) is often used. In this setting, we address the model selection problem of the regression structure assumed for the outcome variable.

Y. Ninomiya (✉)
The Institute of Statistical Mathematics, Tokyo 190-8562, Japan
e-mail: ninomiya@ism.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_28

For this basic model selection problem, no reasonable information criterion has existed for a long time. Specifically, Platt et al. [9] pioneered the weighted quasi-likelihood information criterion (QICw) and proposed using the quasi-log-likelihood weighted by the inverse of propensity scores as the goodness-of-fit term. While that was quite reasonable, QICw used the number of parameters in the regression structure as the penalty term, which considerably underestimates the bias of the goodness-of-fit term. By asymptotically evaluating the bias and significantly modifying the penalty term, Baba et al. [4] derived the inverse-probability-weighted information criterion (IPWIC). Moreover, Baba and Ninomiya [3] proposed the doubly robust information criterion as a valid one in the setting of doubly robust estimation.

In this paper, we pay attention to the fact that doubly robustness in causal inference does not mean that it is robust against outliers. Then, using a loss function that is robust to outliers, we develop an information criterion that is triply robust in total. As a means of achieving modeling robustness, we use covariate balancing by Imai and Ratkovic [7], rather than the most standard method by Scharfstein et al. [12]. The basis of these ideas is also given by Ninomiya [8], and this paper is its generalization which includes not only the estimation method but also the setting. The organization of subsequent sections is as follows. In Sect. 28.2, after explaining the model and assumptions to be treated, outlier-resistant inverse-probability-weighted estimation and covariate balancing are given. Although covariate balancing does not necessarily suppose that the propensity score estimator converges to the true value, we confirm that the estimation for the parameter of interest tends to be valid. In Sect. 28.3, we propose a risk function for the above setting and give an evaluation of the asymptotic bias for an information criterion as the main result. Based on this, we propose what can be called a generalized triply robust information criterion (gTRIC). Comparisons among QICw, IPWIC, and gTRIC through numerical experiments are performed in Sect. 28.4, and conclusions are summarized in Sect. 28.5.

28.2 Estimation

28.2.1 Model and Assumption

As one of the simplest causal inference models, let us consider

$$y = \sum_{h=1}^{H} t^{(h)} y^{(h)}, \qquad y^{(h)} = \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta} + \varepsilon^{(h)}, \qquad \varepsilon^{(h)} \sim (0, \sigma^2). \tag{28.1}$$

Here, $t^{(h)}$ ($\in \{0, 1\}$) is an assignment variable that becomes 1 when the h-th treatment is assigned ($\sum_{h=1}^{H} t^{(h)} = 1$), $y^{(h)}$ ($\in \mathbb{R}$) is an outcome variable when the h-th treatment is assigned, $\boldsymbol{x}^{(h)}$ ($\in \mathbb{R}^p$) is an explanatory variable for $y^{(h)}$, $\boldsymbol{\theta}$ ($\in \mathbb{R}^p$) is a parameter characterizing the regression structure, $\varepsilon^{(h)}$ is an error variable which is not observed ($h \in \{1, \ldots, H\}$). The distribution of the error is unknown, but we

assume that its expectation is 0, its variance is σ^2, and it is uncorrelated with $\boldsymbol{x}^{(h)}$. Note that y on the left-hand side is an observed outcome variable. While it may be natural to set $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}^{(h)}$ as the regression structure for $y^{(h)}$, note also that it can be written as $\boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}$ by setting $\boldsymbol{x}^{(h)} = \boldsymbol{x} \times \mathbf{e}^{(h)}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)\mathrm{T}}, \ldots, \boldsymbol{\theta}^{(H)\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{e}^{(h)}$ is the unit vector with the h-th element 1.

In the model in (28.1), the $(H-1)$ latent outcome variables $y^{(h)}$ with $t^{(h)} = 0$ are missing, and since $\mathrm{E}(y^{(h)}) \neq \mathrm{E}(y^{(h)} \mid t^{(h)} = 1)$ in general, a naive estimation for $\boldsymbol{\theta}$ would result in bias. Therefore, we suppose that the confounding variable $\boldsymbol{z}$ ($\in \mathbb{R}^q$) between $y^{(h)}$ and $t^{(h)}$ is observed so that the bias can be removed. Although not essential, $\mathrm{E}(\boldsymbol{z}) = \mathbf{0}_q$ is assumed because it leads to a simple expression of the result. We also assume the ignorability condition

$$y^{(h)} \perp\!\!\!\perp t^{(h)} \mid (\boldsymbol{z}, \boldsymbol{x}^{(h)}), \qquad h \in \{1, \ldots, H\}.$$

Since $\boldsymbol{z}$ is called a confounding variable, we suppose that $\mathrm{P}(t^{(h)} = 1 \mid \boldsymbol{z}, \boldsymbol{x}^{(h)}) = \mathrm{P}(t^{(h)} = 1 \mid \boldsymbol{z})$, but $\boldsymbol{x}^{(h)}$ may contain elements in $\boldsymbol{z}$. From these, it also holds that $(y^{(h)}, \boldsymbol{x}^{(h)}) \perp\!\!\!\perp t^{(h)} \mid \boldsymbol{z}$. Moreover, the positivity condition $\mathrm{P}(t^{(h)} = 1 \mid \boldsymbol{z}) > 0$ is assumed. There are N independent samples according to the model in (28.1), and we put subscript $_i$ for the variables of the i-th sample.

28.2.2 *Inverse-Probability-Weighted Estimation Robust to Outliers*

In causal inference, doubly robust estimation is often discussed, where the robustness is against model misspecification. On the other hand, it is not robust against outliers, since the loss is basically the squared loss or negative log-likelihood. Therefore, instead of the squared loss function $(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta})^2$, consider a loss function $\xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta})$ that is robust against outliers. Then, let $\boldsymbol{\theta}^*$ satisfying

$$\mathrm{E}\left\{ \sum_{h=1}^{H} \frac{\partial}{\partial \boldsymbol{\theta}} \xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\} = \mathbf{0}_p \tag{28.2}$$

be the target of estimation. An example of sophisticated $\xi(\cdot)$ is the one based on β-divergence by Basu et al. [5].

Although the target of estimation is defined by (28.2), $y^{(h)}$ for h with $t^{(h)} = 0$ is not observed, and it is inappropriate to consider the loss $\xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta})$ only for what has been observed. On the other hand, if the relationship between $y^{(h)}$ and $\boldsymbol{z}$ can be modeled correctly, for example, if we have

$$\mathrm{E}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \,\middle|\, \boldsymbol{z} \right\} = \boldsymbol{\zeta}^{(h)}(\boldsymbol{z}) \tag{28.3}$$

using a function $\boldsymbol{\zeta}^{(h)}(\cdot)$ $(\in \mathbb{R}^p)$ and $\mathrm{E}\{\boldsymbol{\zeta}^{(h)}(z)\} = \mathbf{0}_p$ holds, then $\boldsymbol{\theta}^*$ can be consistently estimated from the estimating equation based on (28.3). In causal inference, the correct modeling through $\boldsymbol{\zeta}^{(h)}(z)$, whose argument may be high-dimensional, is often difficult, and a semiparametric approach based on the propensity score $e^{(h)}(z;\boldsymbol{\alpha})$ that does not necessarily require that correct modeling is often used. Here, $e^{(h)}(z;\boldsymbol{\alpha})$ is a model of $\mathrm{P}(t^{(h)} = 1 \mid z)$ and $\boldsymbol{\alpha}$ $(\in \mathbb{R}^r)$ is a parameter characterizing that function. In other words, we consider

$$t^{(1)}, \ldots, t^{(H)} \mid z \sim \text{Multinomial}(1; e^{(1)}(z;\boldsymbol{\alpha}), \ldots, e^{(H)}(z;\boldsymbol{\alpha})) \tag{28.4}$$

as a model for the assignment variable. Then, supposing that the estimator $\hat{\boldsymbol{\alpha}}$ is obtained, we denote the limit of $\hat{\boldsymbol{\alpha}}$ by $\boldsymbol{\alpha}^*$. If the modeling by $e^{(h)}(z;\boldsymbol{\alpha})$ is correct, then $\boldsymbol{\alpha}^*$ is the true value. Among the semiparametric approaches, let us consider the inverse-probability-weighted estimation by Robins et al. [10]. In this approach, the observed values are multiplied by the inverse of the propensity scores as weights to pseudo-recover missing variables, and then the usual estimation is performed. Namely, the estimator $\hat{\boldsymbol{\theta}}$ is given by solving

$$\sum_{i=1}^{N}\sum_{h=1}^{H} \frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})} \frac{\partial}{\partial \boldsymbol{\theta}} \xi(y_i^{(h)} - \boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}_p. \tag{28.5}$$

28.2.3 Covariate Balancing

As for $\hat{\boldsymbol{\alpha}}$ appearing in Sect. 28.2.2, the maximum likelihood estimator is often used, but we would trust the modeling for the assignment variable too much compared to the semiparametric modeling for the outcome variable. Therefore, we do not trust the form of $e^{(h)}(z;\boldsymbol{\alpha})$ as much and use also the information of $\boldsymbol{\zeta}^{(h)}(z)$ in Sect. 28.2.2, which is called covariate balancing by Imai and Ratkovic [7]. For example, if we assume a linear function of z as $\boldsymbol{\zeta}^{(h)}(z)$, that is, if there is a $p \times q$ matrix $\boldsymbol{B}^{(h)}$ such that $\boldsymbol{\zeta}^{(h)}(z) = \boldsymbol{B}^{(h)}z$, then $\hat{\boldsymbol{\alpha}}$ is given by

$$\operatorname*{argmin}_{\boldsymbol{\alpha}} \sum_{h=1}^{H} \left\| \sum_{i=1}^{N} \frac{t_i^{(h)}}{e^{(h)}(z_i;\boldsymbol{\alpha})} z_i \right\|^2. \tag{28.6}$$

If $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(1)\mathrm{T}}, \ldots, \boldsymbol{\alpha}^{(H)\mathrm{T}})^{\mathrm{T}}$ is so-called just-identified, for example, if $e^{(h)}(z_i;\boldsymbol{\alpha})$ is a function of $z_i^{\mathrm{T}}\boldsymbol{\alpha}^{(h)}$, (28.6) leads to

$$\sum_{i=1}^{N} \frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})} z_i = \mathbf{0}_q, \qquad h \in \{1, \ldots, H\}. \tag{28.7}$$

On the other hand, if $\boldsymbol{\alpha}$ is over-identified, for example, if the dimension of $\boldsymbol{\alpha}$ is not enough to make (28.7) valid, the left-hand side of (28.7) is only close to $\mathbf{0}_q$.

When the model $e^{(h)}(z;\boldsymbol{\alpha})$ is correctly-specified, $\hat{\boldsymbol{\alpha}}$ is consistent and $\boldsymbol{\alpha}^*$ is the true value. Then it holds

$$\mathrm{E}\left[\frac{t_i^{(h)}}{e^{(h)}(z_i;\boldsymbol{\alpha}^*)}\frac{\partial}{\partial\boldsymbol{\theta}}\xi(y_i^{(h)}-\boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]=\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\xi(y_i^{(h)}-\boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]=\mathbf{0}_p. \tag{28.8}$$

By combining (28.8) with (28.5), we obtain $\hat{\boldsymbol{\theta}}\xrightarrow{\mathrm{p}}\boldsymbol{\theta}^*$. On the other hand, even when the model $e^{(h)}(z;\boldsymbol{\alpha})$ is misspecified, if $\boldsymbol{\zeta}^{(h)}(z)$ is included in the supposed model, $\hat{\boldsymbol{\theta}}$ should not perform poorly. Actually, if $\boldsymbol{\zeta}^{(h)}(z)=\boldsymbol{B}^{(h)}z$, we have

$$\mathrm{E}\left[\frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})}\frac{\partial}{\partial\boldsymbol{\theta}}\xi(y_i^{(h)}-\boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\,\Big|\,z_i\right]=\boldsymbol{B}^{(h)}\mathrm{E}\left[\frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})}z_i\,\Big|\,z_i\right],$$

and therefore

$$\mathrm{E}\left[\sum_{h=1}^{H}\frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})}\frac{\partial}{\partial\boldsymbol{\theta}}\xi(y_i^{(h)}-\boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]=\sum_{h=1}^{H}\boldsymbol{B}^{(h)}\mathrm{E}\left[\frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})}z_i\right]=\mathbf{0}_p \tag{28.9}$$

holds from (28.7). By combining (28.9) with (28.5), we obtain $\hat{\boldsymbol{\theta}}\xrightarrow{\mathrm{p}}\boldsymbol{\theta}^*$. In summary, we obtain the following.

Proposition 1 *Suppose that the model $\boldsymbol{e}^{(h)}(z;\boldsymbol{\alpha})$ for the assignment variable is correct. Or suppose that the model of the outcome variable $\boldsymbol{\zeta}^{(h)}(z)$ is correct and $\boldsymbol{\alpha}$ is just-identified. If $\boldsymbol{\alpha}$ is over-identified, suppose that $\boldsymbol{B}^{(h)}$ is a matrix such that the right-hand side of (28.9) is $\mathbf{0}_p$. Then, under regularity conditions, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}^*$.*

Since (28.6) makes (28.7) hold approximately, even if $\boldsymbol{B}^{(h)}$ does not satisfy the condition in Proposition 1, $\hat{\boldsymbol{\theta}}$ is approximately consistent.

If the consistency is guaranteed, then from the usual Taylor expansion of (28.5) around $\boldsymbol{\theta}^*$, we obtain

$$\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*=\boldsymbol{A}_1(\boldsymbol{\theta}^*,\boldsymbol{\alpha}^*)\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\frac{t_i^{(h)}}{e^{(h)}(z_i;\hat{\boldsymbol{\alpha}})}\frac{\partial}{\partial\boldsymbol{\theta}}\xi(y_i^{(h)}-\boldsymbol{x}_i^{(h)\mathrm{T}}\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\{1+\mathrm{o_P}(1)\}, \tag{28.10}$$

where

$$\boldsymbol{A}_1(\boldsymbol{\theta},\boldsymbol{\alpha})=\mathrm{E}\left[\sum_{h=1}^{H}\frac{t^{(h)}}{e^{(h)}(z;\boldsymbol{\alpha})}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\xi(y^{(h)}-\boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta})\right]^{-1}. \tag{28.11}$$

In addition, the following matrix is defined for later use.

$$\boldsymbol{A}_2(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \mathrm{E}\left[\sum_{h=1}^{H} \frac{t^{(h)}}{e^{(h)}(\boldsymbol{z}; \boldsymbol{\alpha}) e^{(h)}(\boldsymbol{z}; \hat{\boldsymbol{\alpha}})} \frac{\partial}{\partial \boldsymbol{\theta}} \xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \xi(y^{(h)} - \boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}) \right]. \quad (28.12)$$

28.3 Model Selection

For the setting in Sect. 28.2, we derive an AIC-type information criterion for the selection of the regression structure $\boldsymbol{x}^{(h)\mathrm{T}}\boldsymbol{\theta}$. As usual, let $(\tilde{y}_i^{(h)}, \tilde{t}_i^{(h)}, \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i)$ be a copy of $(y_i^{(h)}, t_i^{(h)}, \boldsymbol{x}_i, \boldsymbol{z}_i)$, that is, these random vectors follow the same distribution independently, and consider a risk function

$$\mathrm{E}\left[\sum_{i=1}^{N} \sum_{h=1}^{H} \frac{\tilde{t}_i^{(h)}}{e^{(h)}(\tilde{\boldsymbol{z}}_i; \boldsymbol{\alpha}^*)} \xi(\tilde{y}_i^{(h)} - \tilde{\boldsymbol{x}}_i^{(h)\mathrm{T}}\hat{\boldsymbol{\theta}}) \right] \quad (28.13)$$

based on the loss function used in the estimation. The naive estimator is the statistic obtained by replacing $(\tilde{y}_i^{(h)}, \tilde{t}_i^{(h)}, \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i)$ with $(y_i^{(h)}, t_i^{(h)}, \boldsymbol{x}_i, \boldsymbol{z}_i)$ and removing the expectation in (28.13), but it tends to be smaller than the original value of the risk. Therefore, we consider their difference as the bias. Using the Taylor expansion with respect to $\hat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}^*$ and (28.10), we define

$$\begin{aligned} b^{\mathrm{AE}} = \frac{1}{N} \sum_{i,j=1}^{N} \sum_{h,k=1}^{H} \{ & \eta^{(hk)}(t_i^{(h)}, \boldsymbol{z}_i, y_i^{(h)}, \boldsymbol{x}_i^{(h)}, t_j^{(k)}, \boldsymbol{z}_j, y_j^{(k)}, \boldsymbol{x}_j^{(k)}; \boldsymbol{\theta}^*, \boldsymbol{\alpha}^*, \hat{\boldsymbol{\alpha}}) \\ & - \eta^{(hk)}(\tilde{t}_i^{(h)}, \tilde{\boldsymbol{z}}_i, \tilde{y}_i^{(h)}, \tilde{\boldsymbol{x}}_i^{(h)}, t_j^{(k)}, \boldsymbol{z}_j, y_j^{(k)}, \boldsymbol{x}_j^{(k)}; \boldsymbol{\theta}^*, \boldsymbol{\alpha}^*, \hat{\boldsymbol{\alpha}}) \} \end{aligned} \quad (28.14)$$

as an asymptotic equivalent to the statistic made by removing the expectation from some transformation of the bias, where

$$\begin{aligned} & \eta^{(hk)}(t_i, \boldsymbol{z}_i, y_i, \boldsymbol{x}_i, t_j, \boldsymbol{z}_j, y_j, \boldsymbol{x}_j; \boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) \\ & = \frac{t_i}{e^{(h)}(\boldsymbol{z}_i; \boldsymbol{\alpha})} \frac{t_j}{e^{(k)}(\boldsymbol{z}_j; \hat{\boldsymbol{\alpha}})} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \xi(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) \boldsymbol{A}_1(\boldsymbol{\theta}, \boldsymbol{\alpha}) \frac{\partial}{\partial \boldsymbol{\theta}} \xi(y_j - \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\theta}), \end{aligned}$$

and $\mathrm{E}(b^{\mathrm{AE}})$ is used as the asymptotic bias for the correction. Taking into account that either (28.8) or (28.9) holds, we obtain the following.

Proposition 2 *The asymptotic bias of the information criterion when estimated using loss function $\xi(\cdot)$ and covariate balancing is given by*

$$\mathrm{E}(b^{\mathrm{AE}}) = \mathrm{tr}\{\boldsymbol{A}_1(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*)\boldsymbol{A}_2(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*)\}, \tag{28.15}$$

where $\boldsymbol{A}_1(\boldsymbol{\theta}, \boldsymbol{\alpha})$ and $\boldsymbol{A}_2(\boldsymbol{\theta}, \boldsymbol{\alpha})$ are defined by (28.11) *and* (28.12)*, respectively.*

Although $\boldsymbol{A}_1(\cdot, \cdot)$ and $\boldsymbol{A}_2(\cdot, \cdot)$ are defined with expectations that depend on the true distribution, they can be replaced by expectations from empirical distributions, $\hat{\boldsymbol{A}}_1(\cdot, \cdot)$ and $\hat{\boldsymbol{A}}_2(\cdot, \cdot)$. In addition, since $\boldsymbol{\theta}^*$ and $\boldsymbol{\alpha}^*$ are the limits of the estimators, we can substitute $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$ for them, respectively. This gives a statistic that is asymptotically equivalent to (28.15), and as a result, we propose

$$\mathrm{gTRIC} \equiv \sum_{i=1}^{N}\sum_{h=1}^{H} \frac{t_i^{(h)}}{e^{(h)}(\boldsymbol{z}_i; \hat{\boldsymbol{\alpha}})}\xi(y_i^{(h)} - \boldsymbol{x}_i^{(h)\mathrm{T}}\hat{\boldsymbol{\theta}}) + \mathrm{tr}\{\hat{\boldsymbol{A}}_1(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})\hat{\boldsymbol{A}}_2(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})\} \tag{28.16}$$

as the information criterion. It is robust to outliers because it uses a loss function $\xi(\cdot)$, and it is robust to modeling between the assignment and confounding variables and modeling between the latent and confounding variables, because it is a valid criterion as long as one of them is correct. We call it gTRIC (generalized triply robust information criterion) since it deals with a more generalized setting than Ninomiya [8].

28.4 Numerical Experiments

The performance of gTRIC is examined by numerical experiments. Since the purpose here is to grasp the characteristics of gTRIC compared to QICw by Platt et al. [9] and IPWIC by Baba et al. [4], we do not examine its applicability from an applied viewpoint, but treat only a simple setting. Specifically, we set $H = 2$ and consider

$$\xi(v) = 1 - \exp(-0.2v^2)$$

based on the β-divergence by Basu et al. [5] for a normal distribution with known variance. For the model of the outcome variable, we suppose $\boldsymbol{x}^{(1)} = (\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{x}^{(2)} = (\boldsymbol{0}^{\mathrm{T}}, \boldsymbol{x}^{\mathrm{T}})^{\mathrm{T}}$ in (28.1), and consider four candidates $p = 2$, $p = 4$, $p = 6$ and $p = 8$ to be selected. For the model of the assignment variable, we suppose only $q = 1$ in (28.4) and set

$$e^{(1)}(z; \boldsymbol{\alpha}) = \{1 + \exp(-\alpha_1 - z\alpha_2)\}^{-1}, \qquad e^{(2)}(z; \boldsymbol{\alpha}) = 1 - e^{(1)}(z; \boldsymbol{\alpha}).$$

The actual data are generated from

$$
\begin{aligned}
&x_1, x_2, x_3, x_4, z \sim \text{Uniform}(-1, 1), \qquad \epsilon \sim f, \qquad \{x_1, x_2, x_3, x_4, z, \epsilon\} \perp\!\!\!\perp, \\
&\boldsymbol{x}^{(1)} = (x_1, x_2, 0, 0)^{\mathrm{T}}, \qquad \boldsymbol{x}^{(2)} = (0, 0, x_1, x_2)^{\mathrm{T}}, \qquad \boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, -\theta_1^*, -\theta_2^*)^{\mathrm{T}}, \\
&y^{(1)} = \boldsymbol{x}^{(1)\mathrm{T}}\boldsymbol{\theta}^* + \zeta(z) + \epsilon, \qquad y^{(2)} = \boldsymbol{x}^{(2)\mathrm{T}}\boldsymbol{\theta}^* + \zeta(z) + \epsilon, \\
&(t^{(1)}, t^{(2)}) \mid z \sim \text{Binomial}(1; \{1 + \exp\{\kappa(z)\}^{-1}, \{1 + \exp\{-\kappa(z)\}^{-1})
\end{aligned}
$$

with $p = 4$. For f, we give the standard normal distribution "N" when no outliers are contaminated, and "O" with two outliers 25 and -25 mixed into the standard normal distribution when outliers are contaminated. For $\zeta(z)$ and $\kappa(z)$, z is given when correctly-specified "C" and $2|z| - 1$ when misspecified "M". Strictly speaking, $\zeta(z) = z$ gives only approximate correctness. In all subsequent experiments, the number of iterations is set to 1, 000. When the error variance σ^2 is required, its true value is used.

Table 28.1 checks whether the bias evaluation in gTRIC accurately approximates the actual bias. Since the true structure is in the model with $p = 4$, the asymptotic bias evaluation is valid for models with $p \geq 4$. First, to confirm that there is no problem with the estimation, the difference between the estimator and the true value is given as bias($\hat{\boldsymbol{\theta}}$) which calculates $(\sum_{j\in\{1,2\}} |\hat{\theta}_j - \theta_j^*| + \sum_{j\in\{p/2+1,p/2+2\}} |\hat{\theta}_j + \theta_j^*| + \sum_{j\notin\{1,2,p/2+1,p/2+2\}} |\hat{\theta}_j|)/p$, and its dispersion size is given as sd($\hat{\boldsymbol{\theta}}$) which calculates $\sum_{j=1}^{p} \text{sd}(\hat{\theta}_j)/p$. The dispersion size differs depending on the sample size, but in any case, the estimation bias is negligible. Next, we compare the three values to the right. "Tru" is the true value of the bias represented as (28.14), and "Ref" is the asymptotic bias given in (28.15), evaluated by Monte Carlo method. These use the true values of parameters and cannot be given from data alone. On the other hand, "Pro" is the evaluation of the asymptotic bias actually used in (28.16). Unexpectedly, "Ref" tends to be farther from "Tru" than "Pro", but in any case, "Pro" approximates "Tru" quite accurately.

Table 28.2 compares the performance of gTRIC, IPWIC, and QICw. The main index of comparison is the mean squared error "MSE" for the estimation of regression function, specifically $\sum_{j\in\{1,2\}}(x_j\hat{\theta}_j - x_j\theta_j^*)^2 + \sum_{j\in\{p/2+1,p/2+2\}}(x_j\hat{\theta}_j + x_j\theta_j^*)^2 + \sum_{j\notin\{1,2,p/2+1,p/2+2\}}(x_j\hat{\theta}_j)^2$ is calculated. It also computes selection rates of models with $p = 2, 4, 6, 8$. One may think that 4, the number of non-zero θ_j^*, should be selected, but this is not necessarily true when θ_j^* is small, so the selection rate is just a reference index. In any case, QICw clearly has a larger MSE than gTRIC. Since QICw is not a valid criterion with or without outliers and actually underestimates the bias, the overfitting model with $p > 4$ is selected more often. On the other hand, when there are no outliers, IPWIC tends to be superior to gTRIC, although the difference is not large. This is because gTRIC loses some of its effectiveness in dealing with outliers. However, the performance of IPWIC is significantly worse in the presence of outliers. Clearly, IPWIC is choosing simpler models, sometimes too much so, and giving MSEs larger than even QICw. On the other hand, gTRIC is hardly affected by outliers, as expected.

Table 28.1 Examination of the bias evaluation in gTRIC

f	κ	ζ	n	(θ_1^*, θ_2^*)	p	bias($\hat{\boldsymbol{\theta}}$)	sd($\hat{\boldsymbol{\theta}}$)	Tru	Pro	Ref
N	C	C	100	(1.0, 1.0)	2	0.00322	0.350	3.20	3.12	3.28
					3	0.00260	0.352	4.73	4.63	5.10
					4	0.00606	0.352	6.23	6.06	7.06
N	C	C	200	(0.5, 0.5)	2	0.00095	0.239	2.96	3.06	3.13
					3	0.00268	0.240	4.45	4.58	4.77
					4	0.00026	0.242	6.04	6.06	6.46
N	C	C	200	(1.0, 0.5)	2	0.00454	0.234	2.99	3.11	3.17
					3	0.00690	0.237	4.52	4.63	4.82
					4	0.00552	0.240	6.11	6.12	6.52
N	C	C	200	(1.0, 1.0)	2	0.00104	0.279	3.63	3.64	3.76
					3	0.00001	0.278	5.39	5.43	5.75
					4	0.00128	0.281	7.30	7.18	7.83
N	C	C	300	(0.5, 0.5)	2	0.00114	0.228	3.72	3.63	3.72
					3	0.00168	0.229	5.65	5.42	5.66
					4	0.00197	0.226	7.31	7.20	7.63
N	C	M	200	(1.0, 0.5)	2	0.00500	0.238	2.99	3.05	3.13
					3	0.00387	0.239	4.49	4.55	4.77
					4	0.00215	0.237	5.92	6.04	6.46
N	M	C	200	(1.0, 0.5)	2	0.00351	0.227	2.82	2.79	2.87
					3	0.00301	0.229	4.28	4.16	4.36
					4	0.00136	0.229	5.66	5.53	5.91
O	C	C	100	(1.0, 1.0)	2	0.00161	0.351	3.24	3.12	3.32
					3	0.00050	0.353	4.81	4.62	5.17
					4	0.00192	0.356	6.36	6.10	7.15
O	C	C	200	(0.5, 0.5)	2	0.00071	0.238	3.02	3.08	3.15
					3	0.00230	0.239	4.52	4.60	4.80
					4	0.00288	0.240	6.06	6.12	6.51
O	C	C	200	(1.0, 0.5)	2	0.00067	0.242	3.20	3.09	3.16
					3	0.00116	0.241	4.71	4.59	4.80
					4	0.00054	0.242	6.31	6.07	6.50
O	C	C	200	(1.0, 1.0)	2	0.00288	0.240	3.11	3.08	3.13
					3	0.00084	0.240	4.66	4.59	4.76
					4	0.00229	0.241	6.19	6.08	6.45
O	C	C	300	(0.5, 0.5)	2	0.00003	0.191	2.96	3.08	3.12
					3	0.00017	0.191	4.45	4.60	4.71
					4	0.00019	0.193	6.02	6.10	6.34

Table 28.2 Comparison among gTRIC, IPWIC and QICw

f	κ	ζ	n	(θ_1^*, θ_2^*)		MSE	selection rate
N	C	C	100	(1.0, 1.0)	gTRIC	0.2191	(1.4, 71.7, 16.4, 10.5)
					IPWIC	0.2029	(0.3, 64.0, 35.0, 0.7)
					QICw	0.2636	(0.1, 30.0, 27.5, 42.4)
N	C	C	200	(0.5, 0.5)	gTRIC	0.1107	(10.0, 69.7, 12.3, 8.0)
					IPWIC	0.1028	(2.1, 63.6, 29.5, 4.8)
					QICw	0.1296	(0.4, 32.0, 25.2, 42.4)
N	C	C	200	(1.0, 0.5)	gTRIC	0.1095	(10.4, 68.3, 13.4, 7.9)
					IPWIC	0.0972	(1.9, 64.5, 32.3, 1.3)
					QICw	0.1286	(0.2, 29.9, 25.7, 44.2)
N	C	C	200	(1.0, 1.0)	gTRIC	0.1008	(0.0, 78.9, 12.7, 8.4)
					IPWIC	0.0965	(0.0, 67.9, 31.8, 0.3)
					QICw	0.1296	(0.0, 32.3, 25.2, 42.5)
N	C	C	300	(0.5, 0.5)	gTRIC	0.0695	(3.2, 77.7, 12.2, 6.9)
					IPWIC	0.0691	(0.3, 65.2, 28.4, 6.1)
					QICw	0.0870	(0.0, 31.2, 25.8, 43.0)
N	C	M	200	(1.0, 0.5)	gTRIC	0.1115	(10.4, 68.5, 14.2, 6.9)
					IPWIC	0.0997	(2.7, 65.7, 30.8, 0.8)
					QICw	0.1305	(0.7, 33.1, 21.5, 44.7)
N	M	C	200	(1.0, 0.5)	gTRIC	0.0992	(8.5, 72.4, 11.7, 7.4)
					IPWIC	0.0895	(1.3, 65.1, 31.7, 1.9)
					QICw	0.1174	(0.5, 32.1, 25.0, 42.4)
O	C	C	100	(1.0, 1.0)	gTRIC	0.2367	(1.8, 69.4, 16.2, 12.6)
					IPWIC	0.3982	(38.9, 53.4, 7.3, 0.4)
					QICw	0.3415	(17.8, 25.9, 22.7, 33.6)
O	C	C	200	(0.5, 0.5)	gTRIC	0.1118	(9.5, 70.4, 12.8, 7.3)
					IPWIC	0.1332	(37.5, 51.9, 8.3, 2.3)
					QICw	0.1356	(14.0, 26.8, 21.1, 38.1)
O	C	C	200	(1.0, 0.5)	gTRIC	0.1052	(9.8, 71.1, 11.7, 7.4)
					IPWIC	0.1326	(41.9, 48.9, 8.3, 0.9)
					QICw	0.1301	(16.5, 28.2, 21.9, 33.4)
O	C	C	200	(1.0, 1.0)	gTRIC	0.1025	(0.0, 78.9, 13.3, 7.8)
					IPWIC	0.1594	(12.3, 76.2, 11.4, 0.1)
					QICw	0.1559	(5.9, 31.1, 23.1, 39.9)
O	C	C	300	(0.5, 0.5)	gTRIC	0.0698	(2.9, 77.0, 12.7, 7.4)
					IPWIC	0.0956	(27.4, 58.5, 12.2, 1.9)
					QICw	0.0929	(11.3, 30.0, 24.0, 34.7)

28.5 Conclusion

We have derived a triply robust information criterion in a more generalized setting than Ninomiya [8] in causal inference, where a rigorous reevaluation of AIC can significantly change analysis results compared to a formal use of AIC. Numerical experiments have confirmed that the proposal clearly outperforms the existing information criteria when outliers are actually contaminated. While β-divergence was used in the numerical experiments, γ-divergence by Fujisawa and Eguchi [6] is robust to a high percentage of outliers, and the extension of the proposal for γ-divergence should be significant. In addition, causal inference in econometrics has been developing more rapidly than ever in recent years, and generalizing the proposal to accommodate this development is a major challenge for the future. For example, the semiparametric difference-in-difference method by Abadie [1] is a core of development and is well suited to our method. Its extensions have also attracted much attention, and the development of the proposal incorporating such extensions is an interesting topic.

Acknowledgements The author sincerely thanks the reviewers for their helpful and constructive comments. This research was supported by JSPS Grant-in-Aid for Scientific Research 23H00809 and 23K18471.

References

1. Abadie, A.: Semiparametric difference-in-differences estimators. Rev. Econ. Stud. **72**, 1–19 (2005). https://doi.org/10.1111/0034-6527.00321
2. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) 2nd International Symposium on Information Theory, pp. 716–723. Akademiai Kiado, Budapest (1973)
3. Baba, T., Ninomiya, Y.: Doubly robust criterion for causal inference (2021). arXiv: 2110.14525
4. Baba, T., Kanemori, T., Ninomiya, Y.: A C_p criterion for semiparametric causal inference. Biometrika **104**, 845–861 (2017). https://doi.org/10.1093/biomet/asx054
5. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998). https://doi.org/10.1093/biomet/85.3.549
6. Fujisawa, H., Eguchi, S.: Robust parameter estimation with a small bias against heavy contamination. J. Multivariate Anal. **99**, 2053–2081 (2008). https://doi.org/10.1016/j.jmva.2008.02.004
7. Imai, K., Ratkovic, M.: Covariate balancing propensity score. J. Roy. Stat. Soc. B **76**, 243–263 (2014). https://doi.org/10.1111/rssb.12027
8. Ninomiya, Y.: Triply robust information criterion for propensity score analysis. J. Jpn. Stat. Soc. **51**, 275–294 (in Japanese) (2022). https://doi.org/10.11329/jjssj.51.275
9. Platt, R.W., Brookhart, M.A., Cole, S.R., Westreich, D., Schisterman, E.F.: An information criterion for marginal structural models. Stat. Med. **32**, 1383–1393 (2013). https://doi.org/10.1002/sim.5599
10. Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**, 846–866 (1994). https://doi.org/10.1080/01621459.1994.10476818

11. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974). https://doi.org/10.1037/h0037350
12. Scharfstein, D.O., Rotnitzky, A., Robins, J.M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Am. Stat. Assoc. **94**, 1096–1120 (1999). https://doi.org/10.1080/01621459.1999.10473862

Part V
Decision-Making Theory for Economics

Chapter 29
Research on the Use of Cloud-Based AI for Industry-Specific Utilization of Machine Learning and Decision-Making Frameworks

Shunei Norikumo

Abstract Data analysis using cloud computing is becoming mainstream in many companies. In this paper, I would like to clarify a methodology for strategic cloud implementation, clarify the framework for changing decision-making types, and consider trends in the construction of corporate data analysis infrastructure.

29.1 Introduction

In recent years, there has been a growing trend toward cloud computing, which manages and processes data over the Web, in order to build a stable data infrastructure, operate information systems, and do all this from a cost-conscious perspective. Traditional on-site systems, where you can build your own data infrastructure and manage, operate, and maintain it, have the advantage of using custom environments which help to ensure a high level of security. However, there are operational and technical costs and access issues. Convenience has its downsides. On the other hand, cloud-based systems, which can reduce the system operation burden on the company, allow access to data and systems via the web, support remote working, and are popular as systems because they are extremely convenient for end users. Furthermore, the social role of information systems in developed countries is increasing over time, and it is almost impossible to provide a perfect information system that does not suffer system failures. There is a strong demand for quality assurance and reliability.

Against this backdrop, an increasing number of companies are migrating their information systems to the cloud. Typical cloud migration patterns include moving from traditional on-premises data warehouses to cloud infrastructure, or hybrid architectures that blend on-premises corporate data center resources with public cloud services. Considering the increasing business needs for specialized and advanced data analysis, and data sharing through inter-industry collaboration through digital

S. Norikumo (✉)
Faculty of Economics, Doshisha University, Kyoto, Japan
e-mail: snorikum@mail.doshisha.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_29

supply chains, it can be assumed that the trend toward cloud infrastructure will accelerate further.

29.1.1 Social and Academic Background of This Research

In this research, I would like to focus on companies that utilize cloud services and machine learning, and examine AI data analysis and decision-making. Large-scale digital supply chain (inter-industry collaboration) initiatives are underway in the United States, Europe, and the United States regarding the use of cloud services to build data infrastructure and utilize analysis.

In an academic context, the push toward cloud computing is seen as inevitable, but as an urgent issue, research is being conducted in various fields on how to build a data analysis platform and utilize it in decision-making.

The significance of this research is that we are approaching an era in which data has significant value and power. Against this background, I intend to conduct basic research to consider what kind of management strategies should be developed in the future for industries that are data-weak, that is, industries that generate little data, and companies that have difficulty building data infrastructure.

29.2 Data Analysis and Decision Theory Models

Decision-making theories and methods originated from studies of managerial behavior and have been widely covered in management science. This area has also been developed in many other academic fields such as economics, medicine, information systems, psychology, biology, cognitive science, sociology, and philosophy. The framework of the theoretical model of decision-making is introduced in the following three categories [1].

29.2.1 Three Classifications of Decision Theory Models

(a) **Normative decision model**

Mainly an economic approach, for a choice problem under uncertain conditions, the expected value of the options is calculated (probability p gain x) and the ideal solution with the maximum expected value is found. However, in the 19th century, economists pointed out that in human decision-making, it is more rational to choose to maximize the expected utility $p * u(x)$ based on the subjective value (utility) of x rather than the gain x. Von Neumann and Morgenstern [2] mathematically demonstrated the theory

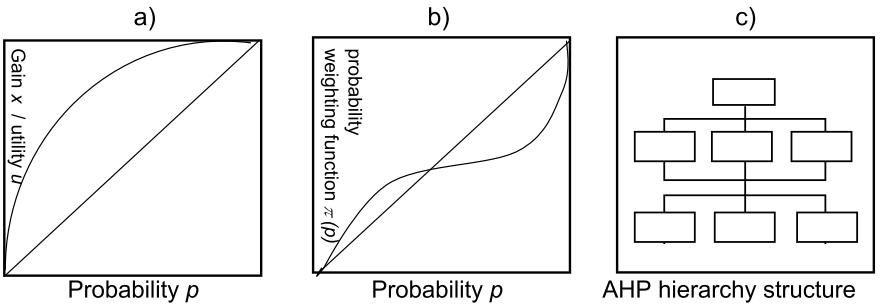


Fig. 29.1 Functions and hierarchy images of each decision-making model

of maximizing expected utility, and it became established as a rational method for decision-making.

(b) **Descriptive decision model**

Mainly a psychological approach, it is a model that describes observed psychological behavior under the assumption that decision makers act according to consistent rules. A typical method is prospect theory, which uses the subjective value $V(X)$ probability weighting function [psychological value] $\pi(p)$ to find a solution. It was devised by behavioral economist Daniel Kahneman (2002 Nobel Prize in Economics) and psychologist Amos Tversky [3] as a model that adds probability weighting functions, framing effects, and loss aversion to expected utility theory.

(c) **Prescriptive decision model**

Mainly using a management/systems approach, choices are supported by systemically analyzing the choice behavior of individuals and groups. In actual choice problems, humans make choices under bounded rationality, and it is difficult to identify and reflect all the factors regarding options and procedures. As a supporting tool, AHP (Analytical Hierarchy Method) [4]was developed by Thomas Sarty (1970s). In addition, a dominant AHP that avoids problems in utility theory has been proposed by Eizo Kinoshita and Masatake Nakanishi [5]. In information systems, the development of methodologies and software (decision support systems) has progressed considerably (Fig. 29.1).

29.2.2 Evolution of Decision-Making Models

Descriptive decision-making models have evolved into a theory that incorporates human characteristics through the fusion of psychology and economics, developing into behavioral economics, not only in terms of the probability of choices but also in human choice behavior. Prescriptive decision-making models have also been devised as highly flexible theories that allow system design to more closely reflect the intentions and psychological factors of humans and users.

29.2.3 From Uniformity to Diversity in Decision-Making and Data Analysis

In today's society, diverse values and opinions are gathered from individuals, and those receiving them are required to share those values. In decision-making, it is important to consider how to integrate and incorporate elements with different values into a single decision-making process. When diversity includes many attributes that are different from other factors, it is thought that the prescriptive decision-making model is more flexible and adaptable to decision makers and selection problems. In the future, the key to new decision-making support tools for decision-making models will be how to reflect the values of the characteristics, and characteristics of diversity as elements and variable data. An easily explainable process is required.

29.2.4 Artificial Intelligence and Decision-Making (Big Data, Cloud and Optimal Data Analysis)

In recent years, mega-tech organizations such as GAFAM in the United States have been using behavioral economics theory, including descriptive decision-making models, to analyze big data. For example, in the ride-sharing industry, surge pricing, prospect theory, and loss aversion. eBay uses framing laws and default effects to build subscription models. Google search engine has auction theory such as second price auction. These technologies are used in combination with artificial intelligence technology to predict sudden changes in demand, cause-and-effect inference, etc., and can process and automate decision-making to instantly change services and prices according to user psychology. This is something that can be easily observed.

29.3 Management Strategies for Cloud AI Migration

There are several approaches to cloud AI migration strategies, and the choice you make will depend on your business needs [6].

- Rehosting

Rehosting is also commonly referred to as lift and shift or the forklift migration approach. It is the simplest type of migration from on-premises to cloud. Applications, data, schemas, and workloads are moved from the data center to an IaaS cloud deployment without any changes. Migration can be easy and smooth, but since the application has not been changed to cloud-native, it limits what you can do with it. Rehosting is a great first migration for organizations new to the cloud, as it provides a low-impact version of processing and workloads on-premises.

- Refactoring

This strategy, also known as rip and replace, or redesign, is much more labor intensive and time consuming than rehosting. This method requires rewriting and rebuilding the application's architecture (and possibly data and schema) before or after the migration. The method is generally carried out after migration. The main benefit of refactoring is that you can essentially redesign your application from the ground up with the cloud in mind and take advantage of the latest and advanced features offered by your cloud service provider (CSP).

Regarding cloud migration costs, initial cloud migration costs are often higher, but in the long run, cloud environments will run more efficiently and practically. For long-term information system management strategies, refactoring is the best option, allowing you to build a high-quality cloud system foundation and demonstrate your strengths.

- Re-platforming

Replatforming, which falls between rehosting and refactoring, involves making some changes to an application while keeping some of its other core elements. For this reason, it is sometimes referred to as moving and improving, or revising. For example, there may be a case where an application changes its usage to interact with a database. Replatforming also works for moving from on-premises infrastructure to IaaS or even PaaS services.

- Replacement

This migration strategy takes data from existing local applications and moves it to a cloud-based software-as-a-service (SaaS) application created by a third party, discarding the original in-house software. This approach makes sense for companies whose applications have been compromised in some way, or who are simply using traditional tools that are considered inferior to third-party SaaS options.

The above is a summary of strategies for migrating information systems from on-site to the cloud. Companies choose to migrate applications and workloads from local data centers to cloud infrastructure in a variety of scenarios. Furthermore, the current appeal of cloud services is that the data analysis tools that accompany them are attractive and allow for seamless data utilization. Organizational leaders are also looking to address specific needs, such as fostering a better and more flexible work environment for DevOps.

29.3.1 Machine Learning Process

Establishing a data analysis platform through the cloud strategy in Chap. 3 will lead to successful data utilization in the next generation. To date, the combination of cloud-based machine learning construction applications and data platforms has been available from various companies for nearly 20 years, and with the introduction of

AutoML, it has become possible to eliminate much of the time-consuming tasks associated with machine learning and so can be completed quickly. Incorporating machine learning into work can bring even greater benefits. AutoML is an acronym for "Automated Machine Learning" and refers to technology that automates some processes in machine learning. AutoML is expected to automate previously manually-processed tasks and reduce the time it takes to build machine learning. A typical example is the deep learning analysis method.

A general data processing model for successful machine learning consists of the following steps.

(1) Define the problem you want to solve, (2) Define the hypothesis, (3) Collect data, (4) Processing, analysis and prediction of data size and orientation, (5) Creation of a machine learning model (rule setting), (6) Operation of the machine learning model, and (7) Application to a business.

AutoML can automate the processes listed above, from "data collection" to "machine learning model creation." Deep learning is a method in which computers themselves find the characteristics that should be learned from data and analyze the data, which reduces human workload and allows data to be analyzed with high accuracy.

There are three advantages to introducing AutoML: These are: (1) Automating machine learning, (2) Speeding up machine learning model creation, and (3) Reducing burdens on data scientists.

29.3.2 Accelerate Machine Learning Model Creation

By introducing AutoML, it is possible to speed up the creation of machine learning models. Creating a machine learning model is complicated and requires many processes, so before AutoML, it took a long time to complete. However, with the advent of AutoML, it has become possible to create machine learning models in a shorter amount of time than before, making it possible to significantly reduce the time it takes to build machine learning. Additionally, AutoML allows you to create machine learning models without any code, so no programming skills are required. A major advantage of AutoML is that it can be used without any special skills.

29.4 Comparison of Machine Learning Using Cloud Services and Introduction by Field

When building a machine learning model, it is necessary to prepare the infrastructure, develop the learning model, train it, deploy it, etc., which is costly when starting from scratch. Cloud-based fully managed services allow the rapid building of a stable machine learning environment. The machine learning operational process includes

instance creation, model construction, training, and deployment in one stop. The disadvantage of using packaged machine learning is that the model is a black box. Unless you are a competent data scientist, it is difficult to explain the results and make adjustments after the fact. The ability to create and build instances and models with no code has made machine learning much less difficult, but this can sometimes appear as a disadvantage [7].

The following is a summary by the author of the representative services provided by cloud AI, mainly machine learning, and the groups of companies that actively utilize them, based on each company's website and company case study [8] special feature.

Regarding the introduction of machine learning at various companies, the use of data in tabular and text formats is becoming established. The use of IBM's Auto AI for bank reception, telephone operator support, chat, etc. is on-track (Table 29.1).

Recently, there has been a high demand for cloud-based AI, and Amazon AWS, Google GCP, etc. are supported by a wide range of users due to their diverse approaches to machine learning and their scalability. Looking at the industry, many large financial institutions have introduced machine learning, and it is of note that companies that are systematically promoting data utilization are introducing it.

29.5 Conclusion

In this article, I have dealt with the recent progress of the transition to cloud services into three categories: methodologies, strategic post-scenarios, and decision-making based on data analysis, as well as the types of cloud services and companies, by sector, that are increasingly adopting them. In the banking and financial sector, data analysis is increasingly being introduced, including cloud service machine learning. The main focus is on analysis using text data. Others in the retail and transportation sectors are using data to optimize their operations. Information, IT, and ICT companies are also using machine learning. In industries where there are many cases of implementation, similar analysis methods are likely to be used with good results, but in industries such as manufacturing, retail, distribution, and education, data analysis is used in niche areas. There are many cases where this is the case, and if it is successful in such cases, it will become a company's strength as a data analysis application that cannot be imitated.

Table 29.1 Typical services that provide machine learning in the cloud and companies that actively utilize them 2024,02 survey within the scope of companies published on web pages [9–13], etc.

Cloud/ Service	[Overseas] Implementing /industries/utilization	[Japan] Implementing /industries/utilization
Amazon AWS [2006] [Tools and data] SageMaker [Features] No code/low code	[finance] Goldman Sachs [Communication] Thomson Reuters [Medical] AstraZeneca	[Bank] Mitsubishi UFJ, Mizuho Financial Group, Sumitomo Mitsui Banking Corporation. [Finance] Nomura Holdings, Japan Exchange. [Insurance] Tokio Marine Holdings, Sumitomo Life Insurance. [Card] JCB, Credit Saison [Communication] NTT Docomo [Medical] MICIN [etc.] Weather news, Demae-kan, Hoshino Resorts [Edu] Life is Tech
Google GCP [2008] Vertex AI AutoML Tables AutoML Vision AutoML VideoAI AutoML Natural Language LaMDA, PaLM2 [Features] Hyperparameter adjustment, model monitoring.	[Finance] Mr. Cooper Group [Manufacturing] [Transportation] Nuro [IT] Modiface, Interactions	[Bank] Minna no Bank, Mizuho Financial Group, Bank of Yokohama, Bank of Mitsubishi UFJ [Card] JCB [Manufacturing] LIXIL, Mitsubishi Heavy Industries, SUBARU, Toppan Printing, Asahi Group Holdings, Coca-Cola Bottlers Japan [Retail] 7-Eleven Japan, Aeon Retail, Nitori, Cainz [Communication] NTT Docomo [Gov.] Kimotsuki Town, Kagoshima Prefecture [IT] Gurunavi, Yahoo Japan, Kakaku.com, :DeNA
Microsoft Azure [2010] Azure Machine Learning ChatGPT, GPT-4 [Features] Responsible AI	[Manufacturing] BMW Group, Nestle, PepsiCo [Transportation] Deutsche Bahn, FedEx	[Bank] Seven Bank [Insurance] Meiji Yasuda Life Insurance [Manufacturing] Toyota, Fujifilm, Kao, Mitsubishi Heavy Industries, Daikin Industries, Lixil, [Transportation]Yamato Transport, Tokyo Metro, Taisei Corporation
IBM Watson AI [2006] Auto AI, NeuNetS	[Bank] Credito Emiliano [Finance] H&R Block, TD Ameritrade [Manufacturing] Chevrolet, The North Face, Campbell Soup Company, Health&Volume [Retail] 1-800-Flowers, Staples , Roztayger, Head Racquet [Department]Macy's, Rare Carat [IT] Autodesk, The Weather Company [Edu] Georgia Tech University [etc.] Masters, Omni Earth	[Bank]Sumitomo Mitsui Banking, Corporation, Mizuho Bank, Post Insurance [Insurance] Mitsui Sumitomo Insurance [Communication] Softbank, JAL [Transportation] JR East [Manufacturing] Panasonic, Toyota [Service] Rakuten, Autobacs Seven
Data Robot [2012] Automated Machine Learning Visual AI Auto DL AI Application	[Gov.] American department of Defense, US Army [Manufacturing] Ford Direct, CAT, Warner Bros, Citi Ventures, [Sports] Dodgers [Edu] Florida International University, Tokio Marine Kiln,	[Finance] SBI Holdings [Communication] NTT Docomo, Softbank [Manufacturing] Hitachi, Yamaha, Yanmar, LION, KIRIN, Calbee, Daihatsu [Transportation] ANA [Retail] Acom

References

1. Norikumo, S.: Operations research society of Japan. In: Spring research presentation proceedings, 1–B–1, p. 30 (2024)
2. Von Neumann, Morgenstern.: Theory of games and economic behavior (1944)
3. Daniel Kahneman, Amos Tversky.: Prospect theory: An analysis of decision under risk. Econometrica **47**(2), (1979)
4. Norikumo, S.: Questionnaire survey and AHP data analysis. Ohmsha (2024)
5. Eizo Kinoshita, Masatake Nakanishi.: Proposal of new AHP model in light of dominant relationship among alternatives. J. Oper. Res. Soc. Jpn **42**(2), (1999)
6. Teradata.: On-premises to cloud migration (2023). https://www.teradata.jp/insights/data-platform/on-premises-to-cloud-migration
7. Cloud Ace.: Machine learning/AI comparison of the three major clouds: AWS, Azure, and GCP. https://cloud-ace.jp/column/detail327/
8. Nikkei Crosstech, Moriyama, T., Risako Kokushi.: Thorough comparison of AWS, Azure, and Google Cloud, releasing 39 survey results all at once (2024)
9. Amazon AWS.: Amazon SageMaker customers. https://aws.amazon.com/jp/sagemaker/customers/
10. Google Cloud.: Google Cloud customers. https://cloud.google.com/customers?hl=en
11. Microsoft Azure.: Azure case studies and customer stories. https://azure.microsoft.com/en-us/resources/customer-stories
12. IBM Watson AI.: new generative AI capabilities. https://www.ibm.com/jp-ja/products/watsonx-assistant/artificial-intelligence
13. Data Robot.: Helping Customers Deliver Measurable Value from their AI Investments. https://www.datarobot.com/customers/page/2/

Chapter 30
Calculations by Several Methods for D-AHP Including Hierarchical Alternatives

Takao Ohya

Abstract We have proposed a super pairwise comparison matrix (SPCM) to express all pairwise comparisons in the evaluation process of the dominant analytic hierarchy process (D-AHP) or the multiple dominant AHP (MDAHP) as a single pairwise comparison matrix. This paper shows the calculations for D-AHP including hierarchical alternatives by the eigenvalue method and the geometric mean method, and with SPCM by the logarithmic least squares method (LLSM) and the Harker method.

30.1 Introduction

In actual decision-making, a decision-maker often has a specific alternative (regulating alternative) in mind and makes an evaluation on the basis of the alternative. This was modeled in D-AHP (the dominant AHP), proposed by Kinoshita and Nakanishi [1].

If there are more than one regulating alternatives and the importance of each criterion is inconsistent, the overall evaluation value may differ for each regulating alternative. As a method of integrating the importance in such cases, CCM (the concurrent convergence method) was proposed. Kinoshita and Sekitani [2] showed the convergence of CCM.

Ohya and Kinoshita [3] proposed the geometric mean multiple dominant AHP (GMMDAHP), which integrates weights by using a geometric mean based on an error model to obtain an overall evaluation value.

Ohya and Kinoshita [4] proposed an SPCM (Super Pairwise Comparison Matrix) to express all pairwise comparisons in the evaluation process of the D-AHP or the multiple dominant AHP (MDAHP) as a single pairwise comparison matrix.

Ohya and Kinoshita [5] showed, by means of a numerical counterexample, that in MDAHP an evaluation value resulting from the application of the logarithmic least squares method (LLSM) to an SPCM does not necessarily coincide with that of

T. Ohya (✉)
Kokushikan University, 4-28-1, Setagaya, Setagaya-ku, Tokyo 154-8515, Japan
e-mail: takaohya@kokushikan.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_30

the evaluation value resulting from the application of the geometric mean multiple dominant AHP (GMMDAHP) to the evaluation value obtained from each pairwise comparison matrix by using the geometric mean method.

Ohya and Kinoshita [6] showed, using the error models, that in D-AHP an evaluation value resulting from the application of the logarithmic least squares method (LLSM) to an SPCM necessarily coincides with that of the evaluation value resulting obtained by using the geometric mean method to each pairwise comparison matrix.

Ohya and Kinoshita [7] showed the treatment of hierarchical criteria in D-AHP with a super pairwise comparison matrix.

SPCM of D-AHP or MDAHP is an incomplete pairwise comparison matrix. Therefore, the LLSM based on an error model or an eigenvalue method such as the Harker method [8] or two-stage method [9] is applicable to the calculation of evaluation values from an SPCM.

Ohya and Kinoshita [10] and Ohya [11, 12] showed calculations of SPCM by each method applicable to an incomplete pairwise comparison matrix for the multiple dominant AHP including hierarchical Criteria.

Ohya [13] shows the calculations of SPCM by LLSM, the Harker method, and ITSM for the multiple dominant AHP including hierarchical criteria.

Ohya [14] shows the calculations for D-AHP including hierarchical criteria by the eigenvalue method and the geometric mean method, and with SPCM by LLSM, the Harker method, and ITSM including hierarchical criteria.

Ohya [15] shows the calculations for MDAHP including hierarchical criteria by the CCM, by the geometric mean MDAHP, and from SPCM by the logarithmic least squares method, the Harker method, and the improved two-stage method.

This paper shows the calculations for D-AHP including hierarchical alternatives by the eigenvalue method and the geometric mean method, and with SPCM by the logarithmic least squares method (LLSM) and the Harker method.

30.2 SPCM

The true absolute importance of alternative $a(a = 1, \ldots, A)$ at criterion $c(c = 1, \ldots, C)$ is v_{ca}. The final purpose of the AHP is to obtain the relative value between alternatives of the overall evaluation value $v_a = \sum_{c=1}^{C} v_{ca}$ of alternative a.

The relative comparison values $r_{c'a'}^{ca}$ of importance v_{ca} of alternative a at criteria c compared with the importance $v_{c'a'}$ of alternative a' in criterion c', are arranged in a (CA × CA) or (AC × AC) matrix. This is proposed as the SPCM $\mathbf{R} = \left(r_{c'a'}^{ca}\right) or \left(r_{a'c'}^{ac}\right)$.

In a (CA × CA) matrix, index of alternative changes first. In a (CA × CA) matrix, SPCM's $(\mathrm{A}(\mathrm{c}-1)+a,\ \mathrm{A}\left(\mathrm{c}'-1\right)+a'\mathrm{)th}$ element is $r_{c'a'}^{ca}$.

In an (AC × AC) matrix, index of criteria changes first. In a (AC × AC) matrix, SPCM's $(C(a-1)+c, C\left(a'-1\right)+\mathrm{c}'\mathrm{)th}$ element is $r_{a'c'}^{ac}$.

In an SPCM, symmetric components have a reciprocal relationship as in pairwise comparison matrices. Diagonal elements are 1 and the following relationships are true:

If $r_{c'a'}^{ca}$ exists, then $r_{ca}^{c'a'}$ exists and

$$r_{ca}^{c'a'} = 1/r_{c'a'}^{ca} \tag{30.1}$$

$$r_{ca}^{ca} = 1 \tag{30.2}$$

SPCM of D-AHP or MDAHP is an incomplete pairwise comparison matrix. Therefore, the LLSM based on an error model or an eigenvalue method such as the Harker method [10] or two-stage method is applicable to the calculation of evaluation values from an SPCM.

30.3 Numerical Example of Using SPCM for Calculation of MDAHP

Let us take as an example the hierarchy shown in Fig. 30.1. Six alternatives from 1 to 6 and three criteria from I to III are assumed, where Alternative 1 is the regulating alternative. Alternative 4 to Alternative 6 are grouped as S. The representative alternative in the hierarchy of alternatives may vary by criterion. For example, the alternative that is expected to have the greatest evaluation value for each criterion may be selected as the representative alternative for that criterion. In this example, Alternative 4, Alternative 5, and Alternative 4 are the representative alternatives for Criterion I, Criterion II, and Criterion III, respectively.

As the result of pairwise comparisons between alternatives at criterion c(c = I, . . . , III), the following pairwise comparison matrices $\mathbf{R}_c^A$, $\mathbf{R}_c^S$, c = $I, \ldots, III$ are obtained

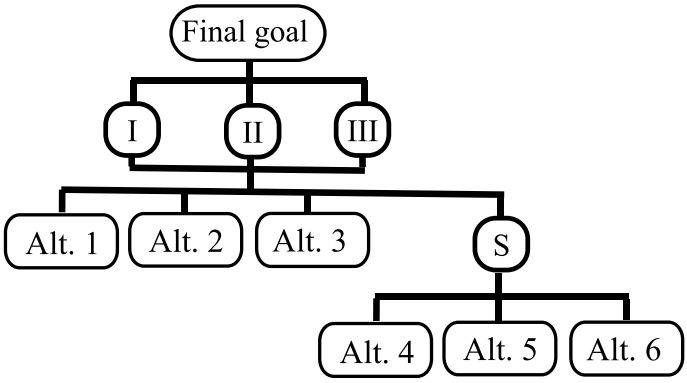


Fig. 30.1 The hieratical structure

$$\mathbf{R}_{\mathrm{I}}^{A} = \begin{matrix} \text{Alt. 1} \\ \text{Alt. 2} \\ \text{Alt. 3} \\ \text{Alt. 4} \end{matrix} \begin{pmatrix} 1 & 1/3 & 3 & 1/2 \\ 3 & 1 & 3 & 1 \\ 1/3 & 1/3 & 1 & 1/3 \\ 2 & 1 & 3 & 1 \end{pmatrix}, \mathbf{R}_{\mathrm{I}}^{S} = \begin{matrix} \text{Alt. 4} \\ \text{Alt. 5} \\ \text{Alt. 6} \end{matrix} \begin{pmatrix} 1 & 1 & 3 \\ 1 & 1 & 2 \\ 1/3 & 1/2 & 1 \end{pmatrix},$$

$$\mathbf{R}_{\mathrm{II}}^{A} = \begin{matrix} \text{Alt. 1} \\ \text{Alt. 2} \\ \text{Alt. 3} \\ \text{Alt. 5} \end{matrix} \begin{pmatrix} 1 & 5 & 3 & 1 \\ 1/5 & 1 & 1/2 & 1/3 \\ 1/3 & 2 & 1 & 1/2 \\ 1 & 3 & 2 & 1 \end{pmatrix}, \mathbf{R}_{\mathrm{II}}^{S} = \begin{matrix} \text{Alt. 4} \\ \text{Alt. 5} \\ \text{Alt. 6} \end{matrix} \begin{pmatrix} 1 & 1/3 & 1/2 \\ 3 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix},$$

$$\mathbf{R}_{\mathrm{III}}^{A} = \begin{matrix} \text{Alt. 1} \\ \text{Alt. 2} \\ \text{Alt. 3} \\ \text{Alt. 4} \end{matrix} \begin{pmatrix} 1 & 1/3 & 1/3 & 1/2 \\ 3 & 1 & 1/3 & 2 \\ 3 & 3 & 1 & 2 \\ 2 & 1/2 & 1/2 & 1 \end{pmatrix}, \mathbf{R}_{\mathrm{III}}^{S} = \begin{matrix} \text{Alt. 4} \\ \text{Alt. 5} \\ \text{Alt. 6} \end{matrix} \begin{pmatrix} 1 & 3 & 2 \\ 1/3 & 1 & 1/2 \\ 1/2 & 2 & 1 \end{pmatrix},$$

With regulating Alternative 1 as the representative alternative, importance between criteria is evaluated by pairwise comparison. As a result, the following pairwise comparison matrices R_{I}^{C} is obtained

$$R_{\mathrm{I}}^{C} = \begin{bmatrix} 1 & 1/2 & 3 \\ 2 & 1 & 3 \\ 1/3 & 1/3 & 1 \end{bmatrix},$$

The (CA × CA) order SPCM for this example is

$R_{(CA\times CA)} =$

	I1	I2	I3	I4	I5	I6	II1	II2	II3	II4	II5	II6	III1	III2	III3	III4	III5	III6
I1	1	1/3	3	1/2			1/2						3					
I2	3	1	3	1														
I3	1/3	1/3	1	1/3														
I4	2	1	3	1	1	3												
I5				1	1	2												
I6				1/3	1/2	1												
II1	2						1	5	3		1		3					
II2							1/5	1	1/2		1/3							
II3							1/3	2	1		1/2							
II4										1	1/3	1/2						
II5							1	3	2	3	1	1						
II6										2	1	1						
III1	1/3						1/3						1	1/3	1/3	1/2		
III2													3	1	1/3	2		
III3													3	3	1	2		
III4													2	1/2	1/2	1	3	2
III5																1/3	1	1/2
III6																1/2	2	1

Table 30.1 Evaluation values obtained by D-AHP with the eigenvalue method

Criterion	I	II	III	Overall evaluation value
Alternative 1	1	1.587	0.420	3.007
Alternative 2	2.060	0.345	1.029	3.433
Alternative 3	0.522	0.613	1.782	2.916
Alternative 4	1.861	0.482	0.727	3.070
Alternative 5	1.626	1.262	0.220	3.109
Alternative 6	0.710	1.103	0.400	2.213

30.4 Results of Calculation of D-AHP by the Eigenvalue Method or the Geometric Mean Method

This section shows the calculations for D-AHP including hierarchical alternative by D-AHP [1] with the eigenvalue method or the geometric mean method.

30.4.1 Results of Calculation of D-AHP by the Eigenvalue Method

Table 30.1 shows the evaluation values obtained by the original D-AHP with the eigenvalue method.

30.4.2 Results of Calculation of D-AHP by the Geometric Mean Method

Table 30.2 shows the evaluation values obtained by the original D-AHP with the geometric mean method.

Table 30.2 Evaluation values obtained by D-AHP with the geometric mean method

Criterion	I	II	III	Overall evaluation value
Alternative 1	1	1.587	0.420	3.007
Alternative 2	2.032	0.344	1.051	3.427
Alternative 3	0.515	0.610	1.851	2.976
Alternative 4	1.789	0.483	0.724	2.996
Alternative 5	1.563	1.266	0.220	3.048
Alternative 6	0.683	1.106	0.398	2.187

30.5 Results of Calculation from SPCM by LLSM

For pairwise comparison values in an SPCM, an error model is assumed as follows:

$$r_{c'a'}^{ca} = \varepsilon_{c'a'}^{ca} \frac{v_{ca}}{v_{c'a'}} \tag{30.3}$$

Taking the logarithms of both sides gives

$$\log r_{c'a'}^{ca} = \log v_{ca} - \log v_{c'a'} + \log \varepsilon_{c'a'}^{ca} \tag{30.4}$$

To simplify the equation, logarithms will be represented by overdots as $\dot{r}_{c'a'}^{ca} = \log r_{c'a'}^{ca}$, $\dot{v}_{ca} = \log v_{ca}$, $\dot{\varepsilon}_{c'a'}^{ca} = \log \varepsilon_{c'a'}^{ca}$. Using this notation, Eq. (30.4) becomes

$$\dot{r}_{c'a'}^{ca} = \dot{v}_{ca} - \dot{v}_{c'a'} + \dot{\varepsilon}_{c'a'}^{ca}, c, c' = 1, ..., C, a, a' = 1, \ldots, A \tag{30.5}$$

From Eqs. (30.1) and (30.2), we have

$$\dot{r}_{c'a'}^{ca} = -\dot{r}_{ca}^{c'a'} \tag{30.6}$$

$$\dot{r}_{ca}^{ca} = 0 \tag{30.7}$$

If $\varepsilon_{c'a'}^{ca}$ is assumed to follow an independent probability distribution of mean 0 and variance σ^2, irrespective of c, a, c', a', the least squares estimate gives the best estimate for the error model of Eq. (30.5) according to the Gauss–Markov theorem.

Equation (30.5) comes to the following Eq. (30.8) by vector notation.

$$\dot{\mathbf{Y}} = \mathbf{S}\dot{\mathbf{x}} + \dot{\boldsymbol{\varepsilon}} \tag{30.8}$$

where

Table 30.3 Evaluation values obtained from SPCM by LLSM

Criterion	I	II	III	Overall evaluation value
Alternative 1	1	1.587	0.420	3.007
Alternative 2	2.060	0.345	1.029	3.433
Alternative 3	0.522	0.613	1.782	2.916
Alternative 4	1.861	0.482	0.727	3.070
Alternative 5	1.626	1.262	0.220	3.109
Alternative 6	0.710	1.103	0.400	2.213

$$\dot{\mathbf{x}} = (\dot{x}_{\mathrm{I2}}\ \dot{x}_{\mathrm{I3}}\ \dot{x}_{\mathrm{I4}}\ \dot{x}_{\mathrm{I5}}\ \dot{x}_{\mathrm{I6}}\ \dot{x}_{\mathrm{II1}}\ \dot{x}_{\mathrm{II2}}\ \dot{x}_{\mathrm{II3}}\ \dot{x}_{\mathrm{II4}}\ \cdots\ \dot{x}_{\mathrm{III5}}\dot{x}_{\mathrm{III6}})^{\mathrm{T}}$$

$$\dot{\mathbf{Y}} = \begin{bmatrix} \dot{r}_{\mathrm{I2}}^{\mathrm{I1}} \\ \dot{r}_{\mathrm{I3}}^{\mathrm{I1}} \\ \dot{r}_{\mathrm{I4}}^{\mathrm{I1}} \\ \dot{r}_{\mathrm{II1}}^{\mathrm{I1}} \\ \dot{r}_{\mathrm{III1}}^{\mathrm{I1}} \\ \dot{r}_{\mathrm{I3}}^{\mathrm{I2}} \\ \dot{r}_{\mathrm{I4}}^{\mathrm{I2}} \\ \dot{r}_{\mathrm{I4}}^{\mathrm{I3}} \\ \dot{r}_{\mathrm{I5}}^{\mathrm{I4}} \\ \dot{r}_{\mathrm{I6}}^{\mathrm{I4}} \\ \dot{r}_{\mathrm{I6}}^{\mathrm{I5}} \\ \dot{r}_{\mathrm{II2}}^{\mathrm{II1}} \\ \vdots \\ \dot{r}_{\mathrm{III6}}^{\mathrm{III5}} \end{bmatrix} = \begin{bmatrix} \log(1/3) \\ \log 3 \\ \log(1/2) \\ \log(1/2) \\ \log 3 \\ \log 3 \\ \log 1 \\ \log(1/3) \\ \log 1 \\ \log 3 \\ \log 2 \\ \log 5 \\ \vdots \\ \log(1/2) \end{bmatrix}, \mathbf{S} = \begin{bmatrix} -1 & & & & & & & & & \cdots & & \\ & -1 & & & & & & & & \cdots & & \\ & & -1 & & & & & & & \cdots & & \\ & & & & & -1 & & & & \cdots & & \\ & & & & & & & & & \cdots & & \\ 1 & -1 & & & & & & & & \cdots & & \\ 1 & & -1 & & & & & & & \cdots & & \\ & 1 & -1 & & & & & & & \cdots & & \\ & & 1 & -1 & & & & & & \cdots & & \\ & & 1 & & -1 & & & & & \cdots & & \\ & & & 1 & -1 & & & & & \cdots & & \\ & & & & & 1 & -1 & & & \cdots & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & & & & & & & & 1 & -1 \end{bmatrix}$$

To simplify calculations, $v_{11} = 1$, that is $\dot{v}_{11} = 0$. The least squares estimates for Formula (30.8) are calculated by $\hat{\dot{\mathbf{x}}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\dot{\mathbf{Y}}$.

Table 30.3 shows the evaluation values obtained from SPCM by LLSM for this example.

30.6 Results of Calculation from SPCM by the Harker Method

In the Harker method, the value of a diagonal element is set to the number of missing entries in the row plus 1 and then evaluation values are obtained by the usual eigenvalue method.

The SPCM by the Harker method for this example is

Table 30.4 Evaluation values obtained from SPCM by the Harker method

Criterion	I	II	III	Overall evaluation value
Alternative 1	1	1.418	0.436	2.854
Alternative 2	2.025	0.281	1.123	3.430
Alternative 3	0.512	0.498	1.997	3.008
Alternative 4	1.729	0.359	0.750	2.839
Alternative 5	1.411	1.013	0.210	2.634
Alternative 6	0.612	0.827	0.380	1.819

$H_{(CA\times CA)} =$

	I1	I2	I3	I4	I5	I6	II1	II2	II3	II4	II5	II6	III1	III2	III3	III4	III5	III6
I1	13	1/3	3	1/2			1/2						3					
I2	3	15	3	1														
I3	1/3	1/3	15	1/3														
I4	2	1	3	13	1	3												
I5				1	16	2												
I6				1/3	1/2	16												
II1	2						13	5	3		1		3					
II2							1/5	15	1/2		1/3							
II3							1/3	2	15		1/2							
II4										16	1/3	1/2						
II5							1	3	2	3	13	1						
II6										2	1	16						
III1	1/3						1/3						13	1/3	1/3	1/2		
III2													3	15	/3	2		
III3													3	3	15	2		
III4													2	1/2	1/2	13	3	2
III5																1/3	16	1/2
III6																1/2	2	16

Table 30.4 shows the evaluation values obtained from the SPCM by the Harker method for this example.

30.7 Conclusion

SPCM of D-AHP is an incomplete pairwise comparison matrix. Therefore, the LLSM based on an error model or an eigenvalue method such as the Harker method or two-stage method is applicable to the calculation of evaluation values from an SPCM.

This paper shows the calculations for D-AHP including hierarchical alternatives by the eigenvalue method and the geometric mean method, and with SPCM by the logarithmic least squares method (LLSM) and the Harker method.

The representative alternative in the hierarchy of alternatives may vary by criterion. For example, the alternative that is expected to have the greatest evaluation value for each criterion may be selected as the representative alternative for that criterion.

References

1. Kinoshita, E., Nakanishi, M.: Proposal of new AHP model in light of dominative relationship among alternatives. J. Oper. Res. Soc. Jpn. **42**, 180–198 (1999)
2. Kinoshita, E., Sekitani, K., Shi, J.: Mathematical properties of dominant AHP and concurrent convergence method. J. Oper. Res. Soc. Jpn. **45**, 198–213 (2002)
3. Ohya, T., Kinoshita, E.: The geometric mean concurrent convergence method. In: Proceedings of the 10th International Symposium on the Analytic Hierarchy Process (2009)
4. Ohya, T., Kinoshita, E.: Proposal of super pairwise comparison matrix. In: Watada, J., et al. (eds.) Intelligent Decision Technologies, pp. 247–254. Springer, Berlin Heidelberg (2011)
5. Ohya, T., Kinoshita, E.: Super pairwise comparison matrix in the multiple dominant AHP. In: Watada, J., et al. (eds.) Intelligent Decision Technologies, Smart Innovation, Systems and Technologies, vol. 1, pp. 319–327. Springer, Berlin, Heidelberg (2012)
6. Ohya, T., Kinoshita, E.: Super pairwise comparison matrix with the logarithmic least squares method. In: Neves-Silva, R., et al. (eds.) Intelligent Decision Technologies, Frontiers in Artificial Intelligence and Applications, vol. 255, pp. 390–398. IOS press (2013)
7. Ohya, T., Kinoshita, E.: The treatment of hierarchical criteria in dominant AHP with super pairwise comparison matrix. In: Neves-Silva, R., et al. (eds.) Smart Digital Futures 2014, pp. 142–148. IOS press (2014)
8. Harker, P.T.: Incomplete pairwise comparisons in the analytic hierarchy process. Math. Model. **9**, 837–848 (1987)
9. Nishizawa, K.: Estimation of unknown comparisons in incomplete AHP and It's compensation. Report of the Research Institute of Industrial Technology, Nihon University Number 77, 10 pp. (2004)
10. Ohya, T., Kinoshita, E.: Super pairwise comparison matrix in the multiple dominant AHP with hierarchical criteria. In: Czarnowski, I., et al. (eds.) KES-IDT 2018, SIST 97, pp. 166–172. Springer International Publishing AG (2019)
11. Ohya, T.: SPCM with Harker method for MDAHP including hierarchical criteria. In: Czarnowski, I., et al. (eds.) Intelligent Decision Technologies 2019, SIST 143, pp. 277–283. Springer International Publishing AG (2019)
12. Ohya, T.: SPCM with improved two stage method for MDAHP including hierarchical criteria. In: Czarnowski, I., et al. (eds.) Intelligent Decision Technologies 2020, SIST 193, pp. 517–523. Springer International Publishing AG (2020)
13. Ohya, T.: Calculations of SPCM by several methods for MDAHP including hierarchical criteria. In: Czarnowski, I., et al. (eds.) Intelligent Decision Technologies 2021, SIST 238, pp. 609–616. Springer International Publishing AG (2021)
14. Ohya, T.: Calculations by several methods for D-AHP including hierarchical criteria. In: Czarnowski, I., et al. (eds.) Intelligent Decision Technologies 2022, SIST 309, pp. 385–393. Springer Nature, Singapore (2022)
15. Ohya, T.: Calculations by several methods for MDAHP including hierarchical criteria. In: Czarnowski, I., et al. (eds.) Intelligent Decision Technologies 2023, SIST 352, pp. 263–272. Springer Nature, Singapore (2023)

Chapter 31
Evaluating Information by Using Unification Among Feature Structures

Takafumi Mizuno

Abstract Evaluation methods among goods or services are unsuitable for evaluating pieces of information due to a property they can copy without cost. In this article, I provide a model to evaluate information. It is based on the assumption that valuable information means that we can add other information to it without inconsistency. Pieces of information are represented as feature structures used in the linguistic processing areas, and adding pieces of information is operated as unification. A simple way to rank information, provided in this article, is ordered by the number of successful unifications among valuable information as criteria.

31.1 Introduction

There are various definitions of "information". In this article, we consider information to be an economic good. Examples include knowledge, technology, know-how, ideas, forecasts, predictions, product descriptions and recommendations, news and gossip, music, novels, movies, and data from comics. Particularly in recent years, software, databases, ICT services, and similar technologies have expanded in society, all of which are also considered as information. The characteristic of such information as a commodity is in its replicability [1]. Economics has focused on exchangeability, and substitution rate has been essential to determine the evaluation of goods. It evaluates many goods and selects the best one. When evaluation becomes challenging, setting criteria for evaluation and conducting pairwise comparisons are also commonly employed. The emphasis on exchangeability is based on the principle that valuable things are scarce.

However, in present economies where information predominates, evaluations based on substitution rates are becoming inappropriate. The information does not have scarcity. Billions of people can access and simultaneously utilize few pieces of information.

T. Mizuno (✉)
Meijo University, 4-102-9 Yada-Minami, Higashi-ku, Nagoya-shi, Aichi, Japan
e-mail: tmizuno@meijo-u.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_31

Today, the value sought in information can be encapsulated as resilience. The ideal scenario is to continue development by accepting various evaluations, thus enabling improvements to be made to a more robust form even in the event of accidents.

In this article, we focus on the monotonicity of information and propose a model for evaluating information based on an operation between pieces of unification: unification.

31.2 Feature Structures and Unification

There are many styles of commodities of information that are copiable at no cost if there are no legal regulations. If additional functions are needed, they can be easily added. It means that we hard to evaluate information by presuming exchange. This article takes the stand that a piece of information is valuable when another valuable piece of information can be added to it without inconsistency. To formalize it, in this section, I introduce a feature structure representing pieces of information and unification adding pieces of information. They have been used in natural language processing and symbolic artificial intelligence research.

A feature structure represents a set of attributes and their values as below.

$$a_1 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \end{bmatrix}. \tag{31.1}$$

It represents a piece of information that is evaluated by city-office, and is verified, and is updated in 2024.

Values of feature structures can be feature structures such as

$$a_2 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{insurance} & \begin{bmatrix} \textit{liability} & \textit{yes} \end{bmatrix} \\ \textit{update} & \textit{2020} \\ \textit{verification} & \textit{no} \end{bmatrix}. \tag{31.2}$$

The value of the attribute *insurance* of the feature structure a_2 is nested feature structure [*liability yes*].

Variables are also feature structures. They represent that there is no information. In this article, a variable denoted by a symbol starts with '?' such as ?X.

And inconsistency is also a feature structure, too.

Unification is an operation between two feature structures to construct a new feature structure that contains all of both information but no additional information. For example, let us consider a unification between the previous feature structure a_1 and c as below.

$$c = \begin{bmatrix} \textit{evaluator} & \text{?X} \\ \textit{verification} & \textit{yes} \end{bmatrix}. \tag{31.3}$$

In this article, the operation is denoted as $a_1 \wedge c$.

$$a_1 \wedge c = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \end{bmatrix}. \tag{31.4}$$

The variable ?X, which has no information, is replaced to *city-office* by adding information.

While the unification between a_2 and c is inconsistent because values of the attribute *verrification* conflict. In this article, a unification is referred to as success when the unification does not make inconsistency.

31.3 An Example of Ranking Pieces of Information

This article assumes that valuable information can add another valuable information to it without inconsistency. That can be modeled as a feature structure evaluated by how many successful unifications are between it and valuable feature structures. The valuable feature structure is referred to as the criteria in this article.

For example, let us consider ranking the following three pieces of information.

$$a_1 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \end{bmatrix}, \tag{31.5}$$

$$a_2 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{insurance} & \begin{bmatrix} \textit{liability} & \textit{yes} \end{bmatrix} \\ \textit{update} & \textit{2020} \\ \textit{verification} & \textit{no} \end{bmatrix}, \tag{31.6}$$

$$a_3 = \begin{bmatrix} \textit{evaluator} & \textit{insurance-company} \\ \textit{insurance} & \begin{bmatrix} \textit{fire} & \textit{yes} \\ \textit{leak} & \textit{no} \end{bmatrix} \\ \textit{update} & \textit{2020} \end{bmatrix}. \tag{31.7}$$

And let us consider evaluating them on three criteria.

$$c_1 = \begin{bmatrix} \textit{update} & \textit{2024} \end{bmatrix}, \tag{31.8}$$

$$c_2 = \begin{bmatrix} \textit{verification} & \textit{yes} \\ \textit{evaluator} & \text{?X} \end{bmatrix}, \tag{31.9}$$

$$c_3 = \begin{bmatrix} \textit{insurance} & \text{?X} \\ \textit{verification} & \textit{yes} \end{bmatrix}. \tag{31.10}$$

Evaluating a_1 is done by unifications among it and criteria as follows.

$$a_1 \wedge c_1 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \end{bmatrix}, \tag{31.11}$$

$$a_1 \wedge c_2 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \end{bmatrix}, \tag{31.12}$$

$$a_1 \wedge c_3 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & \textit{yes} \\ \textit{update} & \textit{2024} \\ \textit{insurance} & \text{?X} \end{bmatrix}. \tag{31.13}$$

Notice that the unification ignores what the information means. In $a_1 \wedge c_3$, information of insurance just added to a_1. It succeeded just because a_1 does not have the information. Valuable information only means that it can add other information without inconsistency.

The feature structure a_2 cannot unify among any criteria; $a_2 \wedge c_1 = a_2 \wedge c_2 = a_2 \wedge c_3 =$ inconsistency. And, unifications of the feature structure a_3 are

$$a_3 \wedge c_1 = \text{inconsistency}, \tag{31.14}$$

$$a_3 \wedge c_2 = \begin{bmatrix} \textit{evaluator} & \textit{insurance company} \\ \textit{insurance} & \begin{bmatrix} \textit{fire} & \textit{yes} \\ \textit{leak} & \textit{no} \end{bmatrix} \\ \textit{update} & \textit{2020} \\ \textit{verification} & \textit{yes} \end{bmatrix}, \tag{31.15}$$

$$a_3 \wedge c_3 = \begin{bmatrix} \textit{evaluator} & \textit{insurance company} \\ \textit{insurance} & \begin{bmatrix} \textit{fire} & \textit{yes} \\ \textit{leak} & \textit{no} \end{bmatrix} \\ \textit{update} & \textit{2020} \\ \textit{verification} & \textit{yes} \end{bmatrix}. \tag{31.16}$$

The number of succeeded unifications of a_1, a_2, and a_3 are 3, 0, and 2. We can rank them $a_1 \succ a_3 \succ a_2$.

31.3.1 How to Find Criteria

The above example assumes that we understand which pieces of information are valuable, which are referred to as criteria. In actual cases, however, it is not easy to prepare them. If we obtain information as a set of feature structures, we can prepare criteria as common structures that almost all have. An operation that extracts common information between two feature structures is a generalization.

For example, the generalization between previous a_1 and a_2 be

$$a_1 \vee a_2 = \begin{bmatrix} \textit{evaluator} & \textit{city-office} \\ \textit{verification} & ?\mathrm{X} \\ \textit{update} & ?\mathrm{Y} \end{bmatrix}. \tag{31.17}$$

The operation finds a feature structure that contains only the common information in both. If values conflict, they are replaced with variables; if an attribute is not on one side, the attribute is deleted.

31.4 Conclusions

Recent information and communication technology advances have clarified criteria for evaluating information as a commodity, such as monotonicity and resilience. These are characteristics that have not been emphasized in previous services.

The property of information not being tampered with is referred to as monotonicity. When monotonicity is preserved, while editing or deleting information is permissible, records of these operations persist, and the information only increases monotonically. The reason value emerges from monotonicity is because information inherently lacks it. In other words, information is highly susceptible to tampering. In blockchain, an infrastructure for giving values to information, hash functions play a crucial role in ensuring monotonicity [2]. The model of evaluating information provided in this article focuses on monotonicity.

In this article, the value of information is whether we can add more valuable information to it monotonically. To model and verify this, we used unifications among feature structures. They have been used in linguistics processing and logic programming. We can ignore computational costs when we use the approach only for decision-making or ranking. The actual usage feature structure may be bigger and more complex. We must define the operation and provide efficient procedures to construct an application that treats such a vast structure. The efficient unification algorithms already have been developed [3–6].

Information is a commodity that can be duplicated and provided at zero marginal cost. Generally, such commodities introduce significant asymmetry within the macro economy. In a typical economy, transactions occur through exchangings. Currency is exchanged as compensation for receiving goods or labor. In such transactions, the provider loses goods or labor. However, when currency is exchanged as compensation for receiving information, the provider still retains the information and additionally gains currency. Companies selling information receive currency as compensation for providing duplication. Unlike companies selling goods or services, companies treat information as commodity transfer duplications, leading to a monotonic asset increase. There will be a need to reconsider the currency mechanism in the future. The approach of this study is expected to be helpful as a mechanism for evaluating information.

References

1. Noguchi, Y.: Economic Theory of Information (in Japanese). Toyo-Keizai-Shinpo-sha, Japna (1974)
2. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System (2024). https://bitcoin.org/bitcoin.pdf. Accessed 17 Feb 2024
3. Pollard, C., Sag, I.: INFORMATION-BASED THEORY OF SYNTAX AND SEMANTICS, vol. 1, The Center for the Study of Language and Information Publications (1987)
4. Pereira, F.: A structure-sharing representation for unification-based grammar formalisms. In: 23rd Annual Meeting of the Association for Computational Linguistics, pp. 137–144 (1985)
5. Wroblewski, D.: Nondestructive graph unification. Proc. AAAI Conf. Artif. Intell. **6**, 582–587 (1987)
6. Tomabechi, H.: Quasi-destructive graph unification with structure-sharing, COLING '92: Proceedings of the 14th Conference on Computational Linguistics, vol. 2, pp. 440–446 (1992)

Part VI
Large-Scale Systems for Intelligent Decision-Making and Knowledge Engineering

Chapter 32
Crisisology-Based Decision-Making Model in Updating Geographical Information Systems for Regions

Boris Ulitin, Eduard Babkin, and Sergey V. Zykov

Abstract Geographic information systems are a universal source of up-to-date information about the region. Moreover, this information is heterogeneous (from spatial data to descriptive data) and is used when making decisions by all possible stakeholders: economic partners, enterprises, other regions, e-government. In this regard, the urgent task is to make a decision on regular updating of the GIS, based on various factors (from technical to economic, socio-cultural and organizational). To automate and simplify this task we propose a method and a corresponding decision support software service for linguistic multi-criteria choice among multiple design alternatives. Our method and prototype are based on a hierarchy of cross-disciplinary criteria and ontology-based mechanisms (for the dynamic GUI customization) which reflect the concept of crisisology.

32.1 Introduction

Over the past decades, production in general and software engineering in particular were understood and practiced in different ways. Changeable business constraints, complex technical requirements, and the so-called human factors imposed on the software solutions caused what was articulated as "software crises", which typically result from an imbalance between available resources, business requirements, and technical constraints [1]. These complex sources of trouble require a multifaceted approach (as well as a related software) to address each of their layers.

Geographic information systems (GISs) demonstrate a bright example of such an information system, that allow for the acquisition, processing, and sharing of spatial

B. Ulitin · E. Babkin · S. V. Zykov (✉)
HSE University, Moscow, Russia
e-mail: szykov@hse.ru

B. Ulitin
e-mail: bulitin@hse.ru

E. Babkin
e-mail: eababkin@hse.ru

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_32

data and related descriptive information about objects within the coverage area of the system. Its primary function is to present spatial data in an interactive manner: with a GIS, users can select the specific spatial information they want, and view the related data associated with it [6].

Advanced technologies and applications enable instant access to data, selection, analysis and the production of reports, with the aim of facilitating decision making and allowing the user to select the best solution [10].

GIS store two types of data simultaneously, which are known as spatial and descriptive data. Spatial data contain information about the shape and location of objects in a specified reference system (geometric data) and the spatial relationships between objects (topological data), and are presented in the form of digital maps. Descriptive data, on the other hand, are any kind of information that does not have a spatial reference, such as the attributes of objects or phenomena, and are represented in tabular form [10]. From this point of view, GIS is a universal source of information available to companies when making business decisions about further development (opening new offices, entering the market, etc.). What is more important, GISs are essential for making strategic decisions that impact the development of a specific area, and the use of a GIS in the administrative decision-making process can play a significant role in management integration [11].

This is completely consistent with the general idea of the research scheme within the framework of crisisology: (1) identification of key BTH factors (business, technology, human); (2) categorization of factors and establishment of dependencies (category theory); (3) description of factors and dependencies in the object language; (4) primary "rough" optimization (AHP, ACDM/ATAM, etc.) with semi-automatic evaluation and ranking of alternatives; (5) secondary "fine" optimization (if required); (6) immersion in an applied object environment based on a virtual machine (VM based on category/combinator theory); (7) final search for the optimal solution (DSS).

However, the most important limitation for the application of this scheme is the fact that the proposed assessments of various factors by experts (and represented in GISs) are often not formalized, but in the form of linguistic information that requires further processing.

In this paper, we will focus on the definition of the problem, decision-making goals and criteria used to update GISs in different regions. We also propose a software service that allows the users processing such assessments in a linguistic form to make a multi-criteria decision on the update of GISs.

This article presents our results as follows. Section 32.2 is devoted to the description of the proposed method for linguistic multi-criteria choice and proposed hierarchy of criteria. In Sect. 32.3 we observe the problem of updating GIS for regions and highlight the main criteria for decision-making. Sections 32.4 and 32.5 contain information on the proposed decision support service design and its usage, as well as the results of evaluating its quality and efficiency. We conclude the article with an analysis of results presented and further research steps.

The research is supported by grant of the Russian Science Foundation (project № 23-21-00112 "Models and methods to support sustainable development of socio-technical systems in digital transformation under crisis conditions").

32.2 A Proposed Method for Hierarchical Multi-criteria Choice

Since in details the proposed multilevel multi-attribute linguistic decision making (ML–MA–LDM) approach was described in [7], here we focus only on the essential stages in terms of the solving task of GIS update decision making.

There are numerous attempts to elaborate new decision-making approaches or adopt existing ones to real-life cases, like healthcare [2], performance evaluation of partnerships, fiber composites optimization, reverse logistics evaluation [3], project resources scheduling [4], supplier selection, aircraft incident analysis [5]. Usually traditional approaches like TOPSIS [2], ELECTRE, VIKOR [3] are used.

On the other hand, in many cases, information that comes from the experts is heterogeneous due to its multigranularity and there are approaches (and methods) to work with such information: the fusion approach for managing multigranular linguistic information [8], the linguistic hierarchy approach and the method of extended linguistic hierarchies [9].

It is important to emphasize that very few existing approaches focus on both types of estimations. At the same time, modern methodologies are likely to assume that there are a number of experts without capturing the area of their expertise as well as the fact that criteria also belong to different abstraction levels (BTH in our case). More importantly, existing methods for decision making are demonstrated on artificial cases with very few experts and alternative solutions. This brings us to the point to propose a new methodology which could incorporate most of the gaps described above.

The proposed ML–MA–LDM approach consists of several consecutive steps starting from defining the estimation rules and finishing with the communication stage. It is important to note that these steps can be found individually in various papers describing the decision-making process, for example in [8], but never were fused in a consistent way. The proposed approach includes:

1. Setting up rules for providing estimations and distribution of criteria weights.
2. Defining available linguistic sets, a context-free grammar and transformation function;
3. Multi-level definition of the desired state, criteria and alternatives: (a) analyzing the desired state on each level of abstraction; (b) formulating criteria for each level of abstraction; (c) formulating alternatives.
4. Giving multi-level and multi-criteria evaluations: (a) aggregating information; (b) searching for the best alternative; (c) communicating the solution found.

After criteria and alternatives were defined, all experts start giving evaluations of each alternative for each available criterion. Let $x = \{x_1, x_2, \ldots, x_N\}$ is the list of alternatives, $c = \{c_1, c_2, \ldots, c_M\}$ is the list of criteria, $e = \{e_1, e_2, \ldots, e_T\}$ is the list of experts. each expert e_k can evaluate alternatives using different linguistic scales S_{g_k} with granularity g_k. In the case of comparative evaluations, we also have the grammar G_H which can be also used for creation of linguistic evaluations. Moreover,

the criteria are given for each level of abstraction in the meta-decision framework, i.e. let $l = \{l_1, l_2, \ldots, l_Z\}$ be the list of the levels of abstraction. Therefore, one evaluation for each given alternative is obtained and the best alternative can be found by sorting these evaluations according to rules of comparing hesitant 2-tuple fuzzy sets. As a result, for each expert we get a matrix of evaluations $R_k = \left(T_{S_{g_k}}^{ij}\right)_{N \times M}$, where $T_{S_{g_k}}^{ij}$ —an evaluation of the expert e_k for the i-th alternative on the j-th criterion in the format of HFLTS on the scale S_g.

Carrying out successively several aggregation of evaluations for each level of abstraction and transformations of these estimates, described in detail in [17], finally we obtain the total evaluation for each i-th alternative and for each level of abstraction as $T_i = MHTWA_{S_{g_k}}^{q}\left(T_{S_{g_k}}^{i_1}, T_{S_{g_k}}^{i_2}, \ldots, T_{S_{g_k}}^{i_Z}\right)$, where i—the index of alternative; q—the vector of weights of levels of abstraction, $q = (q_1, q_2, \ldots, q_Z)^T$, $q_j \geq 0$, $\sum_{j=1}^{Z} q_j = 1$. So, we get the following vector of evaluations $r = \left(T_{S_{g_k}}^{i}\right)_N$, where $T_{S_{g_k}}^{i}$ is the aggregated evaluation for i-th alternative in a form of HFLTS on the scale S_{g_k}.

As a result, we get assessments that draw insights on how each alternative is measured on each abstraction level, that can be used by a decision maker to better understand the scope of alternatives and their influence on each aspect of the problem situation.

32.3 Problem Formulation

The primary goal of a Geographic Information System (GIS) is to meet the demands of the public for geographical spatial information. This entails addressing the requirements of public authorities as well as non-administrative users, including businesses and individuals. The specific aims encompass optimizing the gathering of spatial information within the administration, enhancing data collection, upkeep, and storage, and improving the accessibility of spatial data for both the administration and other data consumers, and creating new information products using existing data [12].

The availability and utilization of accurate and current spatial information for decision-making is a significant concern for any country (especially within the European Union). Addressing these issues necessitates collaborative efforts to exchange, access, and utilize spatial data and services across different levels of government and sectors of the economy. This is all the more important in connection with the introduction of various directives and legal acts that determine the need to keep information systems up to date, ensuring the possibility of their mutual integration (for example, INSPIRE Directive 2007/2/EC [12], which mandates that Member States develop and implement their own spatial information infrastructure as part of the larger spatial information infrastructure of the European Union [13].

On the other hand, GIS is part of the more global information and communication technology (ICT) of e-government, which refers to the utilization of technology to

increase access to and distribution of government information and services to citizens, business partners, employees, other agencies (including government agencies) [14]. The implementation of e-government calls for a holistic approach that takes into account various dimensions (organizational, economic, cultural, social, political, and technical) and different stages of development (from basic information provision to personalized services) at both the strategic and technical levels [18].

What is more important, the factors that influence the success of e-government adoption refer to the areas and operations that should be given priority in order to achieve the best results from e-government adoption, using the crisisology theory. In what follows we focus only on the one of such factors, for making a decision on updating a GIS of the specific region.

Various criteria for the quality of GIS implementation are mentioned in the literature. From a technical point of view, it is possible to evaluate the relevance of data in a GIS [19]. From an economic point of view, it is possible to evaluate factors such as the scope and options for the GIS infrastructure, the phases and options in terms of the benefits of GIS implementation, the projected benefits and costs, the cost-effectiveness indicators, etc. [20].

However, as described earlier in Sect. 32.2, the proposed approach allows us to combine all these quality criteria within one approach, grouping them into the following criteria categories:

Economic aspects, including the financial situation of the regions, potential economic risks and benefits of GIS modernization, and an overview of public outlays;
Socio-cultural aspects, such as the information culture, the potential exclusion of citizens due to numerous factors, and public demand for e-services;
Technological aspects, especially licensing, standardization, the interoperability and integration of systems, and the quality and maturity of e-services;
Organizational aspects, such as compliance with e-government strategy and the adaptation of new management models [18, 20].

32.4 Description of the Decision Support Service

In order to work with the criteria listed above and evaluate the architecture of IT systems according to them, we use a software prototype that includes a backend responsible for the ranking of alternatives and a frontend (GUI) necessary for setting all the components required for evaluation (hierarchy of criteria, alternatives, assessments of alternatives by experts, etc.).

As mentioned earlier, the first step in deciding to upgrade a GIS is to determine the criteria system. It is important to remember that the criteria, by their nature, can be presented in various forms: numerical, textual (linguistic), etc. At the same time, even numerical criteria may differ in assessment scales in terms of quality (from lower to higher and vice versa). Therefore, in the created service, the first stage is the creation of a system of criteria and the setting of scales for their evaluation.

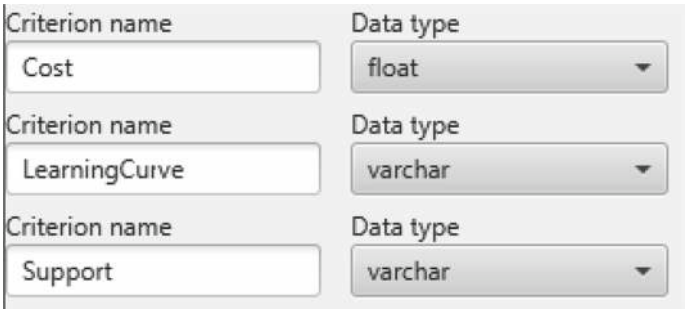


Fig. 32.1 GUI fragments

Using the appropriate GUI (Fig. 32.1), the user is able to first determine and save the name of the criteria system, then the name of each individual criterion, and select the most appropriate data type to represent it: *integer* (int) and *fractional* (float) *numbers*—for quantitative data (e.g. number of users), *date*—in the case of distinguishing newer information systems from older ones, *textual* (varchar)—to represent linguistic assessments. What is more important, we can use not only the data types, but full-fledged domains for each individual evaluation criterion, that is fully consistent with the general theory of categories, the main provisions of which are used in the approach proposed.

A subsequent input of assessments on alternatives by experts is stored in an object-relation database (ORDB) structure, that allows both to prevent the input of contradictory (in terms of type to the corresponding criteria) values, and to save all assessments in the form of a single object (table) for subsequent transfer to the decision-making service.

After the system of criteria and the scale of its evaluation are set, the system switches to the mode of introducing alternatives by experts using the appropriate GUI (Fig. 32.2—top). In this case, the expert has access to only his own estimates for all alternatives, the estimates of other experts are not available to him.

The most significant advantage of the implemented interface is its adaptability. The GUI is generated completely automatically, adjusting to a previously created set of criteria and their data type. If the system of criteria is changed, the GUI will be updated automatically to allow the expert to make assessments in accordance with the updated state of the criteria system.

After all experts have entered their own ratings of alternatives, the entered ratings transfer to the service for comparing and ranking alternatives (see Sect. 32.5). Data transfer in this case is carried out using a JSON package containing all the information necessary for evaluation. As a result of processing the package by the alternative evaluation service, a response JSON packet arrives containing ranked alternatives in descending order of their preference (Fig. 32.2—bottom).

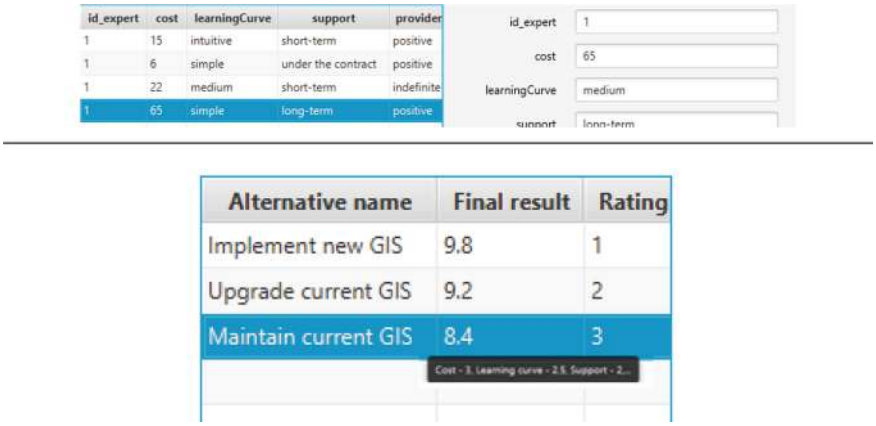


Fig. 32.2 GUI for entering (top) and displaying ranked (bottom) alternatives by the experts

32.5 Case Study of the Proposed Hierarchical Structure of the Criteria and the Prototype

To demonstrate the application of the prototype and the proposed decision model, we consider three main scenarios regarding GIS updating: (1) maintaining the existing GIS without any changes, (2) upgrading the existing GIS using existing technology within existing standards, (3) implementing a completely new GIS using new technology, which renders the previous system obsolete. Discontinuing the usage or terminating the operation of a GIS is not a viable solution, as this software is extensively utilized, particularly in the decision-making procedures of public government entities.

For each criteria group, described in Sect. 32.3, we introduce two-dimensional evaluation. When evaluating the alternatives, experts evaluate, what negative impact will a certain decision have on the different factors included in the criteria: from very low (9) to very high (1).

Each criteria group is divided into individual aspects, based on research in which a group of factors was related to a model of e-government adoption and certain factors were measured in terms of their influence on a scale of one to five, rated by experts [18]:

Economic aspect—*Cost*. This criterion is quantitative, and the evaluation scale is inversely proportional to its values.

Socio-cultural aspect—*Learning curve*. The easier it is to learn how to use, configure, install, and maintain a GIS, the more attractive it is. Therefore, low or shallow learning curves are given higher ratings, and steeper learning curves are given lower ratings.

Socio-cultural aspect—*Support*. Better support is more attractive and is therefore given higher ratings, and weaker support is given lower ratings.

Technological aspect—*Interoperability*. Shows, whether GIS is inoperable with other internal systems and poses data security risks.
Organizational aspect—*Schedule*. Describes the possible total delay in the GIS development and implementation schedule (in days). The greater the possible delay, the less preferred the alternative.
Quality. Serves to assess the impact of a potential decline in quality of GIS on the vital processes within the region. The greater the assessment of a given risk, the less preferred the alternative.

Arbitrarily fractional (counting) scales can be used to evaluate the indicated criteria. Within the framework of this study, we will adhere to the following scales (Table 32.1).

As a result of expert evaluation, the following decision matrix was derived (Table 32.2). Comparing results in terms of different alternatives, we can see logically explained results. For example, results in terms of economic aspects can be interpreted as follows: maintaining current GIS is the cheapest option in this point of time, but it might be more costly in the future, as a result its evaluation is maximum among all alternatives for each expert.

Evaluating the aggregate result (Fig. 32.2—bottom), we find that, taking into account all categories, updating the GIS is optimal, since this option wins according to most criteria (both technological—longer support, interoperability with new systems, and organizational—the new system most likely contains the most up-to-date information, at least from the point of view of the time of creation).

Table 32.1 Criteria evaluation scales

Criterion	Type	Scale (from least preferred to most preferred)
Cost	Quantitative	
Learning curve	Qualitative	None, complex, medium, simple, intuitive
Support		None, under the contract, short-term, long-term
Interoperability		None/does not support current formats/supports current formats (up to 50%)/supports all current formats
Schedule		10 segments: from 1 month to 0 days
Quality		Worse/comparable//higher/is the standard for analogues

Table 32.2 Fragment of a decision matrix, made by 3 experts

Alternative	Economic aspects	Socio-cultural aspects	Technological aspects	Organizational aspects
Maintain current GIS	5,7,9	1,1,3	1,1,3	3,5,7
Upgrade current GIS	3,5,7	1,3,5	1,3,5	5,7,9
Implement new GIS	1,1,3	7,9,9	7,9,9	1,3,5

32.6 Conclusion

In this work, an approach and tools were considered that allow for a multi-criteria assessment of various alternatives based on linguistic data. The proposed approach consists in setting individual rating scales for each criterion and does not require a transition from qualitative to quantitative criteria. The approach is universal and can be applied to various subject areas, which was demonstrated on the example of making a decision on updating GIS for regions.

It is important to note that all manipulations with the criteria and assessments of experts are carried out through a GUI that is adaptive and automatically adjusts to the system of criteria that is set at the initial stage of evaluating alternatives and can be changed in real time. This favorably distinguishes the proposed approach and software prototype from existing systems [15, 16], which are highly specialized and require to unify the used scales for evaluating various criteria in advance and do not allow changing them in the process of working with the system. This makes the system applicable for organizing digital transformation in crisis conditions, since it allows stakeholders to set a complete system of diverse criteria in the original linguistic form, without pre-processing. In the future, we plan to test the prototype on a larger array of data and add support for hierarchical definition of evaluation criteria based on the mechanisms of category theory. This can help to evaluate several decision-making models in parallel, taking into account the varying degrees of influence of factors on the cumulative assessment, making the system more universal and flexible.

References

1. Zykov, S.V.: IT Crisisology: smart Crisis Management in Software Engineering: models, Methods, Patterns, Practices, Case Studies. Springer International Publishing, Switzerland (2021)
2. Dehe, B., Bamford, D.: Development, test and comparison of two multiple criteria decision analysis (MCDA) models: a case of healthcare infrastructure location. Expert Syst. Appl. **42**(19), 6717–6727 (2015)
3. Senthil, S., Srirangacharyulu, B., Ramesh, A.: A robust hybrid multi-criteria decision-making methodology for contractor evaluation and selection in third-party reverse logistics. Expert Syst. Appl. **41**(1), 50–58 (2014)
4. Markou, C., Koulinas, G.K., Vavatsikos, A.P.: Project resources scheduling and leveling using multi-attribute decision models: models implementation and case study. Expert Syst. Appl. **77**, 160–169 (2017)
5. Skorupski, J.: Multi-criteria group decision making under uncertainty with application to air traffic safety. Expert Syst. Appl. **41**(16), 7406–7414 (2014)
6. Ramadhan, A., Sensuse, D., Arymurthy, A.: Assessment of GIS implementation in Indonesian e-Government system. In: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia (2011)
7. Zykov, S.V., Babkin, E., Ulitin, B., Demidovskij, A.: Designing sustainable digitalization: crisisology-based tradeoff optimization in sociotechnical systems. In: Intelligent Decision Technologies. Proceedings of the 15th KES-IDT 2023 Conference, vol. 352, pp. 250–260

8. Herrera, F., Herrera-Viedma, E., Martinez, L.: A fusion approach for managing multi-granularity linguistic term sets in decision making. Fuzzy Sets Syst. **114**(1), 43–58 (2000)
9. Herrera, F., Martinez, L.: A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **31**(2), 227–234 (2001)
10. Monarcha-Matlak, A.: Wykorzystanie systemów informacji przestrzennej w administracji publicznej. In: Szpor G. i Czaplicki K. (eds.) Internet: informacja przestrzenna, pp. 3–12, C.H. Beck, Warsaw (2018)
11. Ganczar, M.: Implementation of the INSPIRE directive into national legal order in the field of infrastructure for spatial information. Studia Prawnicze KUL **3**(83), 91–95 (2020)
12. Izdebski, W.: Spatial information in Poland—Theory and practice. Roczniki Geomatyki XV **2**(77), 175186 (2017)
13. Ogryzek, M., Tarantino, E., Rząsa, K.: Infrastructure of the spatial information in the European community (INSPIRE) based on examples of Italy and Poland. Int. J. Geo-Inf. **9**(12) (2020)
14. Ziemba, E., Papaj, T.: Implementation of e-government in Poland with the example of the Silesian Voivodship. Bus. Inform. **3**(25), 207–221 (2012)
15. Martínez, L., Rodriguez, R.M., Herrera, F.: Linguistic decision making and computing with words. In: The 2-Tuple Linguistic Model. Springer (2014)
16. Wu, J.-T., Wang, J.-Q., Wang, J., Zhang, H.-Y., Chen, X.-H.: Hesitant fuzzy linguistic multicriteria decision-making method based on generalized prioritized aggregation operator. Sci. World J. **2014** (2014)
17. Demidovskij, A., Babkin, E.: Neural multigranular 2-tuple average operator in neural-symbolic decision support systems. In: Proceedings of the Fifth International Scientific Conference "Intelligent Information Technologies for Industry" (2022)
18. Ziemba, E., Papaj, T., Żelazny, R.: A model of success factors for e-government adoption—The case of Poland. Issues Inf. Syst. **14**(2), 87–100 (2013)
19. Choi, T., Chandler, S.: Knowledge vacuum: an organizational learning dynamic of how e-government innovations fail. Gov. Inf. Q. **37**(1) (2020)
20. Ziemba, E., Papaj, T., Żelazny, R., Jadamus-Hacura, M.: Factors influencing the success of e-government. J. Comput. Inf. Syst. **56**(2), 156–167 (2016)

Chapter 33
A Survey of Machine Learning's Integration into Traditional Software Risk Management

Gerald B. Imbugwa, Tom Gilb, and Manuel Mazzara

Abstract Advancements in software development within the era of Industry 4.0 are prompting a reevaluation of traditional risk management methodologies. This survey investigates the integration of machine learning (ML) with established frameworks like those from the Project Management Institute (PMI) and ISO 31000. The study focuses on the potential of ML to enhance risk evaluation and mitigation in the Software Development Life Cycle (SDLC), while also addressing the challenges it brings, such as data privacy concerns and the risk of biased algorithms, especially in dynamic and regulated environments. This research underscores a significant gap in the literature, highlighting the lack of studies specifically focusing on risk assessment in software production using ML, thus positioning itself as a meta-research study. The findings point to a paradigm shift toward an interdisciplinary approach that merges ML with traditional risk management techniques. Despite the complexities and ethical dilemmas introduced by ML, the study emphasizes the dual role of ML in enhancing software quality and introducing intricate challenges, highlighting the need for continuous research and innovation for the effective integration of ML in software risk management.

33.1 Introduction

33.1.1 Background

The landscape of software development risk management is evolving rapidly, with traditional frameworks like the Project Management Institute's (PMI) guidelines and ISO 31000 standards becoming less effective due to the complexity of modern software systems. These traditional methods are increasingly inadequate for addressing

G. B. Imbugwa · T. Gilb (✉) · M. Mazzara
Innopolis University, Russian Federation, Independent Researcher, Oslo, Norway
e-mail: Tom@Gilb.com

G. B. Imbugwa
e-mail: g.imbugwa@innopolis.university

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_33

the needs of critical sectors such as healthcare, finance, and cybersecurity, necessitating more flexible and predictive risk management strategies. Machine learning (ML) offers significant potential by enabling the analysis of large datasets for predictive insights, thus representing a shift in risk management practices [1, 2].

33.1.2 Objectives of the Study

This study aims to perform a Literature Synthesis on the use of ML in software development risk management, focusing on two main objectives:

1. **Review and Analyze ML's Role in Risk Management**: Through a systematic literature review, this study will assess the incorporation of ML techniques, particularly classification, into risk management within software development. It aims to chart the transition from traditional methods to ML-based approaches, highlighting methodologies, impacts, challenges, and successes as reported in the literature [3, 4].
2. **Synthesize Ethical Considerations of ML in Risk Management**: The research will aggregate findings on the ethical implications of using ML in risk management, focusing on data privacy, algorithmic bias, and regulatory compliance. This objective is to provide a comprehensive view of the ethical landscape, underscoring the importance of ethical integrity in ML applications.

Methodological Approach: The study will adhere to a structured framework for Literature Synthesis, with defined search strategies, inclusion and exclusion criteria, and rigorous quality assessment methods to ensure the reliability and validity of findings. The approach will involve data extraction and thematic analysis, synthesizing findings to offer insights into ML's evolution in risk management.

Evaluating the Approach's Accuracy: The accuracy of the Literature Synthesis will be evaluated based on the comprehensiveness of the search strategy, the objectivity of the study selection and quality assessment processes, and the depth of synthesis and critical analysis. Adherence to PRISMA guidelines will serve as a benchmark for methodological rigor.

This focused Literature Synthesis aims to enrich understanding of ML's advancements and ethical considerations in risk management, contributing to informed future research and application in the field.

33.2 Methodology

33.2.1 Database Selection and Exclusion Rationale

To ensure a comprehensive and unbiased literature review, we meticulously selected databases foundational to Computer Science, Machine Learning, and Software

Development research. From an initial list of ten, three were considered based on trial searches assessing volume and relevance:

1. **IEEE Xplore**: Extensive coverage in engineering and technology, emphasizing Machine Learning and Software Development.
2. **ACM Digital Library**: Broad scope in computer science literature, valuable for software development research.
3. **Google Scholar**: Wide-reaching, including diverse fields and publishers, providing comprehensive search results.

Exclusions were made where databases had a limited scope or significant overlap, ensuring focus and efficiency in the search process.

33.2.2 Search Strategy

33.2.2.1 Structured Approach

We employed a structured, replicable search strategy, minimizing bias. A pilot test preceded the main search to refine the strategy, ensuring consistent application across chosen databases.

33.2.2.2 Criteria and Filters

The search covered publications from the last five years to capture recent developments. Inclusion required at least one citation, acknowledging a bias toward established papers but ensuring relevance and recognition.

33.2.2.3 Keyword Selection for Ethical Considerations

To specifically address the study's objective of synthesizing ethical considerations of ML applications in risk management, we included keywords such as "ethical implications", "data privacy", "algorithmic bias", and "regulatory compliance". This ensures that the literature review comprehensively covers the ethical dimensions of ML in risk management.

33.2.2.4 Search Summary

Utilizing Boolean/Phrase mode, the search strategy incorporated "OR" and "AND" operators. The keywords selected reflect the study's focus areas: Machine Learning, Risk Management, and Software Development, with an added emphasis on ethical considerations.

Table 33.1 Predefined inclusion and exclusion criteria

Criteria type	Description
Inclusion	24 See the clean section in Appendix
Exclusion	262 See clean section in Appendix

Table 33.2 Conflict resolution process

Step	Description
Initial review	Two reviewers independently assessed papers
Discussion	Reviewers discuss disagreements
Third reviewer	If no consensus, a third reviewer is consulted

33.2.2.5 Data Export and Reproducibility

"Publish or Perish" was used for Google Scholar, with IEEE Xplore and ACM Digital Library's export procedures detailed in supplementary materials. Our data-cleaning process and paper selection criteria are thoroughly documented to ensure reproducibility, including a specific focus on extracting studies relevant to ethical considerations in ML applications.[1]

33.2.3 Validation Strategy

Two independent reviewers assessed the papers based on predefined inclusion and exclusion criteria, as specified in Supplementary Table 33.1. In cases of disagreements, a consensus was reached through discussion or, if required, a third reviewer would be consulted. The conflict resolution process is further elucidated in Supplementary Table 33.2.

33.2.4 Literature Review

33.2.4.1 Risk Management in Software Production

Foundation Methods

Historically, risk management in software production has been anchored in methodologies such as the Project Management Institute's (PMI) Risk Management Framework and ISO 31000. The PMI Framework, widely recognized in academia and

[1] https://shorturl.at/lmtB8.

industry, offers a structured approach to risk management, emphasizing the identification, assessment, and mitigation of potential pitfalls [1]. In contrast, ISO 31000 provides a more flexible set of guidelines adaptable across various sectors [2]. Tools like SWOT analysis, PERT charts, and FMEA, though pivotal in earlier software development paradigms, are now facing scrutiny with the evolution of software complexity [5]. The System Engineering Architecture (SEA) Planguage methodology presents a robust suite of tools for enhancing Enterprise Risk Management. It emphasizes a comprehensive stakeholder analysis that extends beyond the traditional focus on users and customers, recognizing that overlooked stakeholders pose significant risks [6]. The increasing intricacy of the Software Development Life Cycle (SDLC) necessitates advanced methodologies for early risk prediction [7]. However, Granlund et al. argue that the integration of machine learning in agile software development, particularly in regulated sectors like healthcare, is challenging the efficacy of these foundational methods [8].

Current Challenges and Gaps

The deterministic nature of traditional methodologies is under critical examination in the face of modern software's probabilistic challenges [9]. Hanci et al. note that these traditional methods can fall short in addressing the uncertainties of contemporary software projects, especially in budgeting and effort estimation [10]. Researchers have started looking into new methods and machine learning algorithms, such as ID3 and Naïve Bayes, as possible ways to make the SDLC more predictable, though it's still not clear how well these methods work [10].

33.2.4.2 Machine Learning

Machine Learning and Industry 4.0

The onset of Industry 4.0, characterized by automation, digitalization, and connectivity, has seen a surge in AI implementations. Natural Language Processing (NLP) tools, exemplified by ChatGPT, promise to reshape sectors ranging from customer service to cybersecurity [11]. Di et al. propose AI-based solutions to address these concerns, suggesting automated tools to identify safety issues [12]. Nonetheless, while tools like ChatGPT offer potential, critical voices emphasize the need for rigorous testing and validation before widespread adoption [13].

Machine Learning in Risk Management

Machine learning techniques present a potential avenue to address risk management's modern challenges. Techniques such as Naive Bayesian and Decision Trees have been critically examined for their applicability in software risk assessment [14]. Khan et al.

highlight the emerging trend of Tree-Family Machine Learning (TF-ML) techniques for risk prediction in software requirements [7]. Yet, while these techniques offer promise, Asif et al. argue that the fusion of Case-Based Reasoning (CBR) with machine learning can provide a more nuanced understanding of risk factors [15]. Recent empirical studies emphasize the unique risk factors in ML-based software projects, such as data leakage and overfitting, but consensus on mitigation strategies remains elusive [16].

33.2.4.3 Empirical Insights and Case Studies

Case studies provide in-depth and insightful perspectives on the practical applications of machine learning. Machine learning (ML) has emerged as a significant force in various domains, including medical diagnostics, fintech, and software development. Its impact is characterized by rigorous examination in medical diagnostics, algorithm-driven efficiency in fintech, and simplification of complex tasks in software development.

Medical Analysis

The field of medical analysis has greatly benefited from the implementation of machine learning (ML) techniques. The use of machine learning has resulted in a significant transformation in the field of diagnostic technology. Convolutional neural networks (CNNs), which belong to the domain of deep learning, have significantly advanced the accuracy of X-ray and MRI processing [17, 18].

Fintech

The incisive acumen of ML has revolutionized the financial sector. Here, algorithms crunch vast datasets to unearth patterns for risk assessment, fraud detection, and personalized financial services. The adoption of ML in credit analysis, for instance, has enhanced the ability to capture subtle signals, augmenting the predictive power over traditional creditworthiness measures [19].

ML has transformed fraud detection, which is a crucial concern for financial institutions. By analyzing large datasets, ML algorithms have become adept at identifying fraudulent activities, saving institutions billions of dollars annually [20].

Software Production

Machine learning (ML) has been essential in optimizing the software development life cycle (SDLC), by incorporating intelligent capabilities throughout all stages, from the initial conceptualization to the final deployment. The combination of SDLC

with machine learning operations (MLOps) is exemplified by the framework known as MLASDLC [21].

The efficacy of mentoring in improving the quality of software requirements specifications is demonstrated by a Subject Matter Expert (SME) utilizing the approach outlined by Gilb [6]. This is supported by empirical evidence showing a significant reduction in defect density, emphasizing the importance of mentorship in facilitating the transfer and application of technical knowledge [22].

33.2.4.4 Risks and Challenges

The application of ML is not without its challenges. Data privacy and algorithmic biases are at the forefront of ethical concerns, with scholars and practitioners advocating for transparent and explainable AI models [8]. Implementing ML in regulated domains necessitates not only technical proficiency but also a thorough understanding of legal and ethical constraints [23, 24].

In medical analysis, the stakes are inherently high, with the risks manifesting as potential misdiagnosis and concerns over patient data privacy [25]. Fintech faces similar questions with data security and the imperative to ensure the fairness of algorithms that could profoundly affect financial decisions [26].

Moreover, the potential for bias in ML algorithms looms as a specter across all sectors, leading to discriminatory practices and undermining the integrity of ML applications.

33.2.4.5 Agile Methodologies on Risk Management

Agile frameworks, such as Scrum, provide a degree of flexibility and adaptation; nonetheless, they introduce distinct issues in the domain of risk management, particularly in the context of cybersecurity [27].

In conclusion, there is a growing convergence between conventional risk management practices and machine learning techniques, presenting numerous unexplored opportunities for further investigation. Areas that require further investigation encompass the integration of MLOps practices with stringent industry regulations and the alignment of machine learning (ML) methodologies with Agile approaches [8, 28].

The ethical dimensions of ML in risk management are crucial, given the increasing reliance on algorithmic decision-making. Key ethical principles identified across several AI/ML value statements include design's moral background, expert oversight, and values-driven determinism. These principles emphasize the importance of ethical scrutiny in the design and implementation phases of ML applications, advocating for the construction of systems that are morally sound and accountable [29, 30].

Transparency and interpretability emerge as critical ethical dimensions in ML, with calls for moving beyond black-box models to develop interpretable algorithms from the outset. However, achieving full transparency poses challenges, including

potential privacy breaches, the risk of gaming the system, and the inherent complexity of algorithms. An effective balance between transparency and other ethical considerations, such as accountability and fairness, is, therefore, essential [31–34].

Model risk management (MRM) for AI and ML underscores the need for a governance framework that addresses AI/ML-specific risks throughout the model lifecycle. This framework should ensure that models are conceptually sound, well-controlled, and appropriate for their intended use, enhancing stakeholder trust and accountability [35].

33.3 Discussion

The systematic literature review conducted in this study illuminates a pivotal shift in risk management strategies within the software production domain. Our comprehensive search across three databases yielded 185 articles, yet only 24 of these were found to closely match the theme of integrating machine learning (ML) into risk management within the Software Development Life Cycle (SDLC). Intriguingly, among these, less than five articles specifically addressed the intersection of software production and risk assessment using ML.

This finding underscores the nascent stage of ML application in this field, despite the growing reliance on ML to enhance risk management in line with Industry 4.0's automation, digitalization, and interconnected [11]. Traditional methodologies such as the PMI Framework and ISO 31000 remain foundational but are increasingly complemented by ML-driven approaches that promise heightened adaptability and predictive accuracy amidst escalating software complexity [1, 2].

The limited number of studies directly addressing ML in software risk management also highlights a critical discourse on the practicality and effectiveness of these emerging tools. As noted by Granlund et al., the compatibility of ML with agile development environments is a key concern, suggesting a need for reassessment and potential re-calibration of existing risk frameworks [8].

For practitioners, the integration of ML into risk management offers both opportunities and challenges. While ML can enhance predictability and accuracy in risk assessments, the enduring value of traditional methodologies must not be overlooked, particularly in terms of compliance and ethical considerations [9]. Practitioners are thus advised to judiciously adopt ML, tailoring risk management strategies to specific project needs and constraints.

33.3.1 Limitations

This study acknowledges the limitation in fully encompassing the latest advancements in machine learning (ML) and software development, due to the rapid progress

in these areas. It suggests that future research should emphasize the empirical validation of ML applications in software risk management and assess their actual impact on the industry. The feedback received during the review process has enriched our understanding, enabling us to align our analysis with broader industry trends and the realities faced by professionals in software development. By critically analyzing how risk management in software production is evolving and comparing traditional approaches with ML-based methods, this study provides valuable insights for both researchers and practitioners. It highlights the benefits and challenges of integrating ML into risk management, aiming to facilitate informed decision-making.

Our work contributes to the ongoing dialog on risk management practices in software production, positioning it as a key reference for future research and practical application in the field. It reflects on the transformation of risk management strategies, advocating for a balanced view on the adoption of ML. This study encourages further investigation and collaboration to navigate the complexities of modern software development, ultimately aiming to enhance risk management practices with the aid of ML technologies.

33.4 Conclusion

This systematic literature review analyzes the integration of artificial intelligence (AI), particularly machine learning (ML), into software risk management. It identifies a notable lack of studies on ML-based risk assessment in software production, highlighting an opportunity for future research in this emerging interdisciplinary area.

The study examines the shift from traditional risk management frameworks, such as the PMI Framework and ISO 31000, to ML-driven methods. Traditional approaches remain valuable for managing risks, but the increasing complexity of software systems and the digital environment necessitates more flexible, ML-based solutions. These new methods offer improved insights by analyzing large datasets to identify patterns that might be missed by human analysis.

Nevertheless, incorporating ML into risk management faces challenges, including data privacy concerns, algorithmic biases, and the demand for interpretable, trustworthy models, especially in regulated sectors. These issues underscore the importance of transparency and accountability in ML integration.

The success of ML in other areas, such as medical diagnostics and financial services, suggests its potential to revolutionize software risk management. ML can facilitate both minor improvements and major advancements in risk prediction, assessment, and mitigation.

The goal of this study is to encourage continued exploration and integration of ML with traditional risk management practices. Further empirical research and interdisciplinary cooperation are vital to enhance risk management strategies in software development.

As traditional and modern technologies converge, the software industry must proceed with cautious optimism. Combining ML's predictive capabilities with the solid principles of traditional risk management presents a promising approach to achieving a resilient, adaptable, and robust risk management strategy.

In conclusion, this study highlights the convergence of traditional and technological methodologies in software risk management. It calls for a well-informed, comprehensive approach to software development, preparing for both current and future challenges. This emphasizes the importance of meta-research in spotting and addressing research gaps in the fast-evolving sectors of software and AI.

33.5 Supplementary Table

See Tables 33.1 and 33.2.

References

1. Kerzner, H.: Project Management: A Systems Approach to Planning, Scheduling, and Controlling. Wiley (2017)
2. Iso 31000:2018 risk management—guidelines (2018)
3. Dwork, C.: Differential privacy: a survey of results. Theory Appl. Models Comput. (2008)
4. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown (2016)
5. Vose, D.: Risk Analysis: A Quantitative Guide. Wiley (2008)
6. Gilb, T.: Competitive Engineering: a Handbook for Systems Engineering, Requirements Engineering, and Software Engineering using Planguage. Elsevier (2005)
7. Khan, B., Naseem, R., Alam, I., Khan, I., Alasmary, H., Rahman, T.: Analysis of tree-family machine learning techniques for risk prediction in software requirements. IEEE Access **10**, 98220–98231 (2022)
8. Granlund, T., Stirbu, V., Mikkonen, T.: Towards regulatory-compliant mlops: oravizio's journey from a machine learning experiment to a deployed certified medical product. SN Comput. Sci. **2**(5), 342 (2021)
9. Garousi, V., Felderer, M., Mäntylä, M.V.: Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. Inf Softw Technol (2016)
10. Hancı, A.K.: Risk group prediction of software projects using machine learning algorithm. In: 2021 6th International Conference on Computer Science and Engineering (UBMK), pp. 503–505. IEEE (2021)
11. Javaid, M., Haleem, A., Singh, R.P.: A study on chatgpt for industry 4.0: background, potentials, challenges, and eventualities. J. Econ. Technol. (2023)
12. Di Sorbo, A., Zampetti, F., Visaggio, A., Di Penta, M., Panichella, S.: Automated identification and qualitative characterization of safety concerns reported in UAV software platforms. ACM Trans. Softw. Eng. Methodol. **32**(3), 1–37 (2023)
13. Lianfan, W.: Agile design and AI integration: revolutionizing MVP development for superior product design. Int. J. Educ. Hum. **9**(1), 226–230 (2023)
14. Darandale, S., Mehta, R.: Risk assessment and management using machine learning approaches. In: 2022 international conference on applied artificial intelligence and computing (ICAAIC), pp. 663–667. IEEE (2022)

15. Asif, M., Ahmed, J.: A novel case base reasoning and frequent pattern based decision support system for mitigating software risk factors. IEEE Access **8**, 102278–102291 (2020)
16. Huang, S.-J., Lin, C.-T.: Risk model of machine learning based software project development-a multinational empirical study using modified delphi-ahp method. SSRN 4511875
17. IABAC. Machine learning in healthcare: transforming medical diagnostics. IABAC (2023). URL: iabac.org
18. Krishnamoorthy, P., Vengrenyuk, A., Wasielewski, B., Barman, N., Bander, J., Sweeny, J., Baber, U., Dangas, G., Gidwani, U., Syros, G., et al.: Mobile application to optimize care for ST-segment elevation myocardial infarction patients in a large healthcare system, stemicathaid: rationale and design. European Heart J.-Digit. Health **2**(2), 189–201 (2021)
19. Sadok, F.S.H., El Hadi, M., Maknouzi, E.: Artificial intelligence and bank credit analysis: a review. Cogent Econ. Finance **10**(1), 2023262 (2022)
20. Aeologic: the impact of machine learning in the fintech industry. Aeologic (2023). www.aeologic.com
21. Ranawana, R., Karunananda, A.S.: An agile software development life cycle model for machine learning application development. In: 2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), pp. 1–6 (2021)
22. Terzakis, J.: Reducing requirements defect density by using mentoring to supplement training. Int. J. Adv. Intell. Syst. **6**(1 & 2), 2013 (2013)
23. Almada, M.: Regulating machine learning by design. CPI TechREG Chronicle, February (2023)
24. Althar, R.R., Samanta, D., Kaur, Singh, M.D., Lee, H.-N.: Automated risk management based software security vulnerabilities management. IEEE Access **10**, 90597–90608 (2022)
25. Richens, J.G., Lee, C.M., Johri, S.: Improving the accuracy of medical diagnosis with causal machine learning. Nat. Commun. **11**(1), 3923 (2020)
26. Goodell, J.W., Satish Kumar, Lim, W.M., Pattnaik, D.: Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. J. Behav. Exp. Finance **32**, 100577 (2021)
27. Khurana, S.K., Wassay, M.A.: Towards challenges faced in agile risk management practices. In: 2023 International Conference on Inventive Computation Technologies (ICICT), pp. 937–942. IEEE (2023)
28. Pilliang, M., Tjahjono, B., Sejati, P., Akbar, H., Firmansyah, G. et al.: Criticism of the risk management process in scrum methodology. In: 2022 International Conference on Electrical and Information Technology (IEIT), pp. 338–343. IEEE (2022)
29. Greene, D., Hoffmann, A.L., Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning (2019)
30. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
31. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
32. Thimbleby, H.: NHs number open source software: implications for digital health regulation and development. ACM Trans. Comput. Healthcare **3**(4), 1–26 (2022)
33. Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. **20**(3), 973–989 (2018)
34. De Laat, P.B.: Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos. Technol. **31**(4), 525–541 (2018)
35. Agarwala, G., Latorre, A., Raffel, S.: Model Risk Management for AI and Machine Learning, May 2020

Chapter 34
Advanced Engineering School at Innopolis University: A Global Ecosystem for Future Leaders

Manuel Mazzara, Iouri Kotorov, Yuliya Krasylnykova, Nursultan Askarbekuly, Petr Zhdanov, and Evgenii Bobrov

Abstract This paper outlines the establishment and vision of the Advanced Engineering School (AES) at Innopolis University (IU), a leading institution in Information Technology (IT) and Robotics in Tatarstan, Russia. The AES aims to address global challenges in IT education by training over 13,000 highly qualified IT engineers by 2030, focusing on technological sovereignty and collaboration with industry. The paper discusses the university's current status, the structure of AES, its financial model, scientific research programs, educational policies, Technology Transfer Center (TTC), and strategic partnerships. By aligning with global trends and fostering an innovation ecosystem, Innopolis University aspires to become a global leader in IT education, research, and innovation.

34.1 Background

Innopolis University is a young and ambitious university in Tatarstan, Russian Federation, which has a strong focus on education and on fundamental and applied research in IT and Robotics [1]. It is located in a young city named Innopolis near the capital

M. Mazzara (✉) · N. Askarbekuly · P. Zhdanov · E. Bobrov
Innopolis University, Innopolis, Tatarstan, Russia
e-mail: m.mazzara@innopolis.ru

N. Askarbekuly
e-mail: n.askarbekulyd@innopolis.university

P. Zhdanov
e-mail: pe.zhdanov@innopolis.ru

E. Bobrov
e-mail: e.bobrov@innopolis.ru

I. Kotorov (✉)
Université Paul Sabatier, IRIT, Toulouse, France
e-mail: iouri.kotorov@karelia.fi

I. Kotorov · Y. Krasylnykova
Karelia University of Applied Sciences, Joensuu, Finland

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_34

city of the Republic of Tatarstan, Kazan. Innopolis is also a hometown for a wide range of Information and Communication Technology (ICT) companies and the Innopolis Special Economic Zone. Innopolis aims to be the major Russian IT hub [2]. Since its development, the university has been trying to follow the main trends of IT education typically set by the world's leading higher education institutions [3]. One of these trends was and still is internationalization [4]. Since the very foundation, the university aimed to hire international faculty members and attract international students by organizing international conferences and student competitions [5]. Additionally, the university has been developing various international initiatives, including summer schools, exchange programs, internships, etc. [4]. Existing for ten years, the university succeeded in creating an international environment on campus [5].

The university is a research-active institution aiming to reach visibility at the global level within the scientific community and land high on the international rankings. In order to reach these objectives and increase its reputation, the faculty engages in individual and community efforts to publish articles in high-impact venues following rigorous internationally recognized ethical standards [6].

IU was established on December 10, 2012. At the early stage, it operated in Kazan's downtown and moved to Innopolis city from April to September 2015, after the completion of the new campus. The university's foundation was announced in February 2012, when negotiations were held with Carnegie Mellon University on creating an IT personnel training center in Tatarstan [1].

In July 2012, the president of Tatarstan, Rustam Minnikhanov, met with Gil Taran, head of the iCarnegie Global Learning, a subsidiary of Carnegie Mellon University. The parties agreed to create a new IT university. The city was inaugurated in June 2015. In August 2015, the training of students began in the buildings of the university in the city of Innopolis. Innopolis currently ranks 3rd in Russia for the quality of incoming students, according to the final grades of the school state examination [1].

34.2 Innopolis Today

IU is now approaching its twelfth year and has reached a student population of more than 1000 with 400 employees, of which about 25 faculty members and teaching and research staff of up to 200 people. There are plans to reach 2000 students in the coming years. The university is an integral part of the city, which now has two operational technoparks and two more under construction (at the moment of writing). The flow of students between the university and the companies of technoparks is regular and growing, fulfilling what has been, since the beginning, one of the key measures of success: cooperation with industry and supply of highly qualified professionals.

The university has an extensively developed network of international institutions collaborating under different formats: student exchange, Erasmus +, visiting professors, joint PhD supervision, joint projects, and summer internships offline and online [5]. One of the collaborative projects has seen as partners CERN and Newcastle University [7, 8] and collaborative PhD supervisions involve several universities,

including Toulouse, Nice and Nantes in France, the University of Southern Denmark, the University of Messina in Italy and the University of Brno in the Czech Republic. All these activities dramatically supported the internationalization of the project. In turn, the university's growth and internationalization also helped the city's development, bringing professionals, students, and talents from abroad once the city attracted worldwide attention [4].

With the prerogative of a selective education free of charge and constant attention to the internationalization of teaching and research, Innopolis is aiming at exploiting the benefit of the global trends without suffering the risks, such as excessive emphasis on the market aspects of education and the death run to global rankings which would put at risk the fundamental academic values [9, 10].

IU's model did not suffer particularly during the pandemic, as did those of other countries. Not relying financially on the stream of international students, the struggle of the period was moderate and contained thanks to the ability of management and employees, who all put extra effort into keeping things going [11].

34.3 The Advanced Engineering School

Given the international political and economic environment in 2022, characterized by the introduction of sanctions aimed at software and hardware limitations, issues such as import substitution and training of IT engineers have become of great significance [12]. To ensure technological sovereignty, it is necessary to train highly qualified IT engineers—they are the key competitive advantage of any modern business and will allow them to respond to new national and global challenges.

Considering IU's existing experience and the tasks relevant at the national and global levels, the university sets itself the goal of creating an AES in software engineering. As part of the AES, IU sets the primary task of creating a single ecosystem and conditions for a new type of engineering training in IT [13].

The objective is to train, by 2030, over 13000 highly qualified, competitive IT engineers for companies in the real sector of the economy, able to respond to global scientific and technological challenges and ensure the accelerated development and implementation of domestic software and Artificial Intelligence (AI) services. The mission is the formation of a global ecosystem to develop future leaders and to transition to advanced production of software and hardware systems [14], at the same time developing the city toward the concept of *Smart City* with *Smart and Software-defined Buildings* [15].

The creation of the AES now plays a key role in all educational activities of the university, setting a new vector of development aimed at closer cooperation with industrial partners, including networking [16]. By 2030, IU plans to double the number of undergraduate, graduate, and postgraduate program students. The creation of the AES directly correlates with the goals of increasing the number of students. The creation of new laboratories, a unified environment for developing software solutions, and the active participation of industrial customers all serve as

additional incentives for the best applicants to choose IU. The university plans to expand the number of online educational programs with partner universities and large IT companies that have educational programs for employees [12].

Particular attention is paid to the possibility of internships in industrial partner organizations, the development of professional qualities, and the preservation of the industrial orientation of teacher training. Up to 100% of the teaching staff involved in the AES Program is planned to complete advanced training courses and internships in the partner companies. The new personnel development system involves industrial partners, where the main focus is on the young faculty, which will help achieve the goals in the field of personnel policy [17].

In addition to industrial internships, the university also invites industry representatives to participate in project courses by offering cases and problems for the students to work on. As a result, students work in teams and apply their skills in real projects with the partner companies as the client. The project courses take place on both bachelor and master levels. There are of course organizational and communication challenges connected to conducting such project courses and involving industrial partners. However, this approach has several positive side effects. First, students are introduced to industrial partners, which gives them an opportunity to demonstrate their skills and get hired. Second, students gain hands-on experience and expertise in the stack and problems relevant to the partners. Lastly, it serves as an extra layer of relations between the university and industrial partners, which agrees with the overall idea of AES to combine Education and Industry.

The implementation of the AES program will help achieve the university's ambitious goals of implementing breakthrough research and development in frontier, advanced areas, as well as increase the number of its own software solutions, publications, projects, and level of commercialization [12]. Furthermore, it will significantly accelerate the university's development and attract additional resources from partners.

34.4 The Management and Financial Model

The organizational structure will be developed based on the principles of transparency, integrity, open dialog, evidence-based, collaborative decision-making, financial sustainability, flexibility, and responsible distributed implementation [18]. The Head of the AES will carry out operational activities of the AES. The AES includes three main functional sectors (blocks):

- *The education projects sector*: creates and implements the network of educational programs together with industrial partners, involving the best personnel and engineers of the university, the country, and the world with constant mutual synchronization of the educational, innovative, and scientific activities of the university and partners.

- *The scientific research sector*: implementation of breakthrough R&D together with partners, updating of educational programs (in the field of science), development of the R&D block. The main staff consists of engineers- and teachers who are active in scientific activities, thus eliminating the gap between education and science.
- *The software development and design sector*: implementation of the product part of the frontier task and development of products in cooperation with industrial partners, the formation of practical material (cases) for educational programs.

The financial model of the AES will be based on the experience of the university and will combine funding from budgetary and extrabudgetary funds. It is designed using the principles of building sustainability, stability, investment in development, and product development [12, 19].

34.5 Scientific Research and Development Program

A collaborative initiative between industrial partners and the IU team within the broader scope of the AES is strategically geared toward advancing technologies for an automatic code generation system [12]. This program aims to play a pivotal role in maintaining an innovative ecosystem while simultaneously addressing the imperative of nurturing a new generation of engineering professionals capable of tackling non-trivial challenges, using agile, user-oriented, and gamified processes [20, 21], and spearheading the creation of innovative products, solutions, and technologies [22].

To support the overarching objectives of the AES, a TTC will be established. This center will facilitate the development of the automatic code generation system and serve as a vital hub for commercialization endeavors. A digital project commercialization platform will be implemented to guide and propel projects initiated by the AES students from their conceptualization stages to becoming start-ups or spin-offs within the University company framework [23].

The program envisages a multifaceted approach, incorporating specialized training programs for engineering personnel. These programs will equip students with the requisite skills and knowledge, aligning with industry demands and ensuring their preparedness for the dynamic landscape of technological innovation [17]. The TTC will further catalyze an innovation ecosystem, fostering collaboration, idea exchange, and the development of groundbreaking projects [24].

Key to the success of this program is the collaboration framework with industrial partners, where mutual benefits are sought. The framework is designed to facilitate the seamless exchange of ideas, resources, and expertise between students, researchers, and industry stakeholders [25]. The program's success will be measured through key performance indicators, reflecting achievements such as project commercialization, industry partnerships, and the academic accomplishments of participating students [26].

Looking forward, the program envisions a long-term impact, aligning with the broader goals of the AES. This includes contributing significantly to the advancement

of the automatic code generation field, establishing itself as a cornerstone in the innovation landscape, and solidifying its position as a transformative force in the education and development of future engineering leaders [12].

34.6 Educational Policy

The AES program plans to build on existing foundations and create conditions for the development of a fundamentally new type of engineering training in IT, which allows training professionals to transform industries at Innopolis University. By the end of 2022, 4 network, in-depth higher education programs have been developed and implemented by 2023, in cooperation with high-tech partners and leading Russian and foreign resident experts:

- Bachelor's Degree Program in Information Systems Engineering;
- Master of Technology Degree Program in Software Engineering;
- Master of Technology Degree Program in Security Engineering of
- Systems and Networks;
- Master of Technology Degree Program in Artificial Intelligence and Data Engineering.

The key features of the AES program will be:

- Applied intensive 1+1 Master's degree, 3+1 Bachelor's degree, unique network programs created together with industrial partners, having their own educational schools (VK Education, Sber Academy)
- Engineering approach in teaching: practical work in a single information environment (DevOps pipeline), work with real industry cases, joint work of students in cross-functional project teams
- Introduction of advanced disciplines and tools of partners in the educational process, participation of students in specialized events of partners
- 70% of elective disciplines are taught by industry engineers English, as the only language of instruction, contributes to the
- continuous updating of disciplines, considering global trends in the IT field
- Introduction foreign engineers-residents of the Russian Federation into teaching
- Own an educational platform for continuing education and collection of the digital footprint of students.

The educational paradigm for the implementation of Further Professional training within the AES includes two stages [12]. During the first stage, training will be aimed at solving acute problems of filling the e shortage of highly qualified teachers and mentors capable of teaching and mentoring students in the process of teaching students at an advanced engineering school and their internships in high-tech companies [27]. The second stage represents a systematic, clearly built system of advanced training and retraining of practical engineers and elite teachers based on

a narrow specialization of training, building individual learning trajectories with constant updating of educational programs per the industry's request [28].

34.7 Partnerships

IU plans to implement and develop new training programs and areas of research activities created within the framework of the AES in a network format based on the current partners (Russian educational institutions of higher education). The high level of interest of companies in IT engineers and software solutions developed under the AES is confirmed by the level of companies involved in the project. The total amount of extrabudgetary co-financing from partner companies and sponsors already at the stage of formation of this program is more than 2.3 billion rubles. Among the partners are the largest IT companies in Russia and industry leaders (Gazprom, VK, Abars Bank, etc.).

34.8 Future Perspectives and Global Engagement

34.8.1 Global Trends Shaping IU's Future

As IU embarks on the ambitious journey of establishing the AES, it is crucial to consider the evolving global trends that will shape the landscape of education, technology, and research in the coming years [29]. The university's commitment to staying at the forefront of these trends positions it as a key player in the international academic arena [4].

- *Embracing technological sovereignty*: Technological sovereignty gains paramount importance in a world marked by geopolitical shifts and economic uncertainties. The AES's focus on training IT engineers to navigate software and hardware limitations aligns with the broader global trend of nations seeking self-reliance in critical technological domains [30].
- *International collaboration in research and development*: The success of IU's collaborations with prestigious institutions such as CERN, Newcastle University, and several others demonstrates the potential for further international partnerships. The AES can leverage these existing networks to foster collaborative research and development initiatives, creating a platform for shared expertise and resources [5].
- *Innovation in teaching methodologies*: The future of education is evolving rapidly, driven by technological advancements and changing learning preferences. IU's commitment to an applied, industry-centric approach in teaching, including practical work in a single information environment, positions it well to adapt to emerging trends in educational methodologies [31].

34.8.2 Envisioning the Next Decade: The AES's Role

The AES envisions a transformative role in the local context and on a global scale. By 2030, it aims to produce over 13,000 highly qualified IT engineers, contributing to the accelerated development and implementation of Russian software using domestic and AI tools. This mission aligns with the broader goals of preparing professionals to address global scientific and technological challenges.

- *Closer Cooperation with Industry*: The AES program's success hinges on its ability to foster closer cooperation with industrial partners. By doubling the number of students and creating a unified environment for developing software solutions, the university aims to become a hub for industry-driven innovation. The active participation of industrial customers and the integration of online educational programs with major IT companies signal a commitment to practical, real-world relevance [32].
- *Accelerating Development through Internationalization*: IU's internationalization efforts have played a pivotal role in its growth. As the AES program unfolds, the university envisions expanding international collaborations and attracting top talent from around the world. This influx of diverse perspectives will contribute to the richness of the academic environment and strengthen the university's global standing [29].

34.8.3 Ensuring Sustainable Growth: The Management and Financial Model

A robust management and financial model is imperative to sustain the ambitious goals of the AES program. Transparency, collaboration, and financial sustainability will be the cornerstones of the organizational structure [33]. The three functional sectors—Education Projects, Scientific Research, and Software Development—will work synergistically to achieve the program's objectives [34].

- *Building a financially sustainable model*: Drawing from the university's experience, the financial model for the AES will blend funding from budgetary and extra-budgetary sources. This combination aims to ensure stability, encourage investment in development, and support the creation of innovative products. The financial sustainability of the AES program is essential for its long-term success [35].
- *Nurturing an innovation ecosystem*: IU's commitment to the development of technologies for automatic code generation aligns with the broader goal of nurturing an innovation ecosystem. The Technology Transfer Center's digital project commercialization platform will catalyze student projects into start-ups or spin-offs, contributing to the university's reputation as a hub for entrepreneurial endeavors [36].

34.8.4 Shaping the Future of IT Education: Educational Policy

To shape the future of IT education, the AES program strategically employs educational policies designed to make a lasting impact [37]. The program's unique approach includes immersive applied intensive programs, a strong engineering focus, and active collaboration with industry engineers [38]. This carefully crafted educational policy ensures that graduates not only acquire theoretical knowledge but also develop practical, real-world skills that directly align with the evolving needs of the IT industry [39]. The monitoring of progress is also achieved through the use of dedicated AI tools [40] aligned with the educational goals of the program [41].

Moreover, as part of its comprehensive educational policy, the AES Program prioritizes continuous learning and a global perspective [12]. This is achieved by making English the main language of instruction, introducing foreign engineers-residents into the learning environment, and maintaining a dedicated educational platform for continuous learning. These policy-driven initiatives collectively instill a global outlook among students and foster a culture of perpetual learning [42]. This commitment to adaptability, embedded in the educational policy, becomes instrumental in preparing students to navigate the ever-changing landscape of IT with confidence and competence [12].

- *Fostering industry-relevant skills*: The AES's unique educational approach, including applied intensive programs, an engineering focus, and collaboration with industry engineers, ensures that graduates possess skills directly aligned with industry needs. This focus on practical, real-world applications positions IU's graduates as sought-after assets in the job market [43].
- *Continuous learning and global perspective*: By emphasizing English as the only language of instruction, introducing foreign engineers-residents, and maintaining an educational platform for continuing education, the AES Program not only ensures a global perspective but also fosters a culture of continuous learning. This commitment to adaptability will be crucial in preparing students for the ever-evolving landscape of IT [44].

34.9 Strategic Collaborations: Partnerships for the Future

IU recognizes the significance of strategic partnerships in realizing the goals of the AES program. As the university expands its network of collaborations, it seeks to create new training programs and research activities in a network format, building on the success of existing partnerships.

- *Industry Support and Extrabudgetary Funding*: The substantial extrabudgetary co-financing from industry leaders such as Gazprom, VK and Abars Bank underscores

the high level of interest in the AES program. These partnerships contribute to the program's financial sustainability and provide valuable insights, ensuring that the education provided remains relevant to industry demands [45].

- *A Platform for Innovation and Entrepreneurship*: IU envisions partnerships as more than financial support-they are collaborations that catalyze innovation. The Technology Transfer Center's role in implementing a digital project commercialization platform will provide a structured pathway for turning ideas into start-ups, fostering an entrepreneurial spirit among students and faculty [46].

As IU charts its course into the future with the establishment of the AES, the vision is clear: to become a global leader in IT education, research, and innovation [47]. By aligning with global trends, fostering international collaborations, implementing a sustainable financial model, and prioritizing industry relevance in education, IU is poised to shape the future of IT education and contribute significantly to the advancement of technology on a global scale [48]. The journey ahead holds exciting possibilities, and IU is well-positioned to embrace them with determination and foresight [11].

34.10 Conclusion

The establishment of the AES at IU marks a significant milestone in the institution's journey toward becoming a global leader in IT education, research, and innovation. The university's commitment to staying at the forefront of global trends, fostering international collaborations, implementing a sustainable financial model, and prioritizing industry relevance in education positions it strategically for the future. As IU embarks on this ambitious venture, it aligns itself with key global trends, such as the emphasis on technological sovereignty, international collaboration in research and development, and innovation in teaching methodologies. The AES program envisions a transformative role not only in the local context but on a global scale, aiming to produce over 13,000 highly qualified IT engineers by 2030 and contributing to the accelerated development and implementation of Russian software using domestic and Artificial Intelligence tools. The university's focus on closer cooperation with industry, doubling the number of students, creating a unified environment for developing software solutions, and integrating online educational programs with major IT companies signals a commitment to practical, real-world relevance. The AES program's success hinges on its ability to nurture an innovation ecosystem, drawing from the experience of the TTC and its digital project commercialization platform. The management and financial model of the AES, grounded in transparency, collaboration, and financial sustainability, will play a crucial role in ensuring the program's long-term success. The three functional sectors—Education Projects, Scientific Research, and Software Development—will work synergistically to achieve the program's objectives, fostering industry-relevant skills and continuous learning [49]. IU's commitment to strategic partnerships with industry leaders, such as Gazprom,

VK, and Abars Bank, underscores the AES program's high level of interest and support. These partnerships contribute to financial sustainability and provide valuable insights, ensuring that the education provided remains relevant to industry demands. Looking ahead, the university envisions the AES as a catalyst for transforming into a multidimensional center of excellence focused on pure computer science and synergistic domains of knowledge. Concrete low-level goals, including joining the top 100 universities in the Shanghai Ranking, educating outstanding professionals, and aiming at self-sufficiency, provide a roadmap for the implementation of this vision.

As the journey unfolds, IU is well-positioned to embrace the exciting possibilities that lie ahead with determination and foresight. The AES program stands as a testament to the university's commitment to shaping the future of IT education and contributing significantly to the advancement of technology on a global scale [50, 51].

References

1. Kondratyev, D., Tormasov, A., Stanko, T., Jones, R.C., Taran, G.: Innopolis University-a new it resource for Russia. In: 2013 International Conference on Interactive Collaborative Learning (ICL), pp. 841–848 (2013)
2. Yuloskov, A., Bahrami, M.R., Mazzara, M., Imbugwa, G.B., Ndukwe, I., Kotorov, I.: Traffic light algorithms in smart cities: simulation and analysis. In: Barolli, L. (ed.) Advanced Information Networking and Applications. Lecture Notes in Networks and Systems, (Cham), pp. 222–235. Springer International Publishing (2023)
3. Yuloskov, A., Bahrami, M.R., Mazzara, M., Kotorov, I.: Smart cities in Russia: current situation and insights for future development. Future Internet **13**, 252 (2021)
4. Kotorov, I., Krasylnykova, Y., Zhdanov, P., Mazzara, M.: In: Bruel, J.-M., Capozucca, A., Mazzara, M., Meyer, B., Naumchev, A., Sadovykh, A. (eds.), Internationalization strategy of Innopolis University. In: Frontiers in Software Engineering Education. Lecture Notes in Computer Science (Cham), vol. 12271, pp. 327–340. Springer International Publishing (2020)
5. Aslam, H., Naumcheva, M., Zhdanov, P., Kotorov, I., Mazzara, M., Akhmetgaraeva, E., Valiev, R., Krasylnykova, Y.: Perception of the internationalization process by the University Employees: the case study of Innopolis University. In: Auer, M.E., Pachatz, W., Rüütmann, T. (eds.), Learning in the Age of Digital and Green Transition. Lecture Notes in Networks and Systems, (Cham), pp. 873–883, Springer International Publishing (2023)
6. Kotorov, I., Krasylnykova, Y., Zhdanov, P., Mazzara, M., Aslam, H., Akhmetgaraeva, E., Naumcheva, M., Brown, J.A.: Institutional commitment and leadership as prerequisites for successful comprehensive internationalization. In: Succi, G., Ciancarini, P., Kruglov, A. (eds.), Frontiers in Software Engineering, Communications in Computer and Information Science, (Cham), pp. 1–11. Springer International Publishing (2021)
7. Bauer, R., Breitwieser, L., Meglio, A.D., Johard, L., Kaiser, M., Manca, M., Mazzara, M., Rademakers, M., Talanov, F., Tchitchigin, A.D.: The biodynamo project: experience report. Adv. Res. Biol. Inspired Cogn. Arch. pp. 117–125 (2017)
8. Breitwieser, L., Bauer, R., Di Meglio, A., Johard, L., Kaiser, M., Manca, M., Mazzara, M., Rademakers, F., Talanov, M.: In: The Biodynamo Project: Creating a Platform for Large-Scale Reproducible Biological Simulations (2016). arXiv:1608.04967 [cs]
9. Jibeen, T., Khan, M.A.: Internationalization of higher education: potential benefits and costs. Int. J. Eval. Res. Educ. (IJERE) **4**, 196–199 (2015)

10. Swart, K., Bond-Barnard, T., Chugh, R.: Challenges and critical success factors of digital communication, collaboration and knowledge sharing in project management virtual teams: a review. Int. J. Inf. Syst. Project Manag. **10**, 59–75 (2022)
11. Krasylnykova, Y., Kotorov, I., Demel, J., Mazzara, M., Bobrov, E.: Innopolis University: an agile and resilient academic institution navigating the rocky waters of the covid-19 pandemic. In: Jezic, G., Chen-Burger, J., Kusek, M., Sperka, R., Howlett, R.J., Jain, L.C. (eds.), Agents and Multi-agent Systems: Technologies and Applications 2023. Smart Innovation, Systems and Technologies, (Singapore), pp. 383–392. Springer Nature (2023)
12. Diaz Lantada, A.: Engineering education 5.0: continuously evolving engineering education. Int. J. Eng. Educ. **36**, 1814–1832 (2020)
13. Tushar, M., Ladda, M.R., Saraf, A., Mohammad, S.: Artificial intelligence, its impact on higher education. SSRN Electron. J. **6**, 513–518 (2019)
14. Nagaraj, B.K., K. A, S. B. R, A. S, Sachdev, H.K., S.K. N.: The emerging role of artificial intelligence in stem higher education: acritical review. Int. Res. J. Multidiscip. Technovation, pp. 1–19 (2023)
15. Mazzara, M., Afanasyev, I., Sarangi, S.R., Distefano, S., Kumar, V.: A reference architecture for smart and software-defined buildings (2019)
16. Mattila, M., Yrjölä, M., Hautamäki, P.: Digital transformation of business-to-business sales: what needs to be unlearned? J. Personal Selling Sales Manag. **41**, 113–129 (2021)
17. Moroianu, N., Iacob, S.-E., Constantin, A.: Artificial Intelligence in Education: a Systematic Review, pp. 906–921 (2023)
18. Ruben, B.D.: An Overview of the Leadership Competency Framework, pp. 19–28. Emerald Publishing Limited (2019)
19. Elbanna, S., Thanos, I., Jansen, R.: A literature review of the strategic decision-making context: a synthesis of previous mixed findings and an agenda for the way forward. M@n@gement, vol. 23, pp. 42–60 (2020)
20. Pakhtusova, Y., Megha, S., Askarbekuly, N.: A case study on combining agile and user-centered design. In: Frontiers in Software Engineering: First International Conference, ICFSE 2021, Innopolis, Russia, 17–18, 2021, Revised Selected Papers 1, pp. 47–62, Springer (2021)
21. Vertash, V., Aslam, H., Askarbekuly, N., Mazzara, M.: Introducing gamification into agile processes of a game development company. In: 2021 International Conference Nonlinearity, Information and Robotics (NIR), pp. 1–6. IEEE (2021)
22. Kim, K.-J.: Employer strategy and employee job quality: a real options perspective. Acad. Manag. Proc. **2020**, 20707 (2020)
23. Succi, C., Canovi, M.: Soft skills to enhance graduate employability: comparing students and employers perceptions. Stud. Higher Educ. **45**, 1834–1847 (2020)
24. Trotter, H., Huang, C.-W., Czerniewicz, L.: Seeking equity, agility, and sustainability in the provision of emergency remote teaching during the covid-19 pandemic: A center for teaching and learning takes an expanded role. Higher Learn. Res. Commun. **12**, 1–24 (2022)
25. Assbeihat, J.M.: The impact of collaboration among members on team"s performance. Manag. Adm. Sci. Rev. **5** (2016)
26. Kotorov, I., Pérez-Sanagustín, M., Mansilla, F., Krasylnykova, Y., Hadaou, F.T., Broisin, J.: Supporting the monitoring of institutional competency in learning innovation: the prof-xxi tool. In: 2022 XVII Latin American Conference on Learning Technologies (LACLO), (Armenia, Colombia), pp. 01–08. IEEE (2022)
27. Yamaguchi, D., Tezuka, Y., Suzuki, N.: The differences between winners and losers in competition: the relation of cognitive and emotional aspects during a competition to hemodynamic responses. Adap. Hum. Behav. Physiol. **5**, 31–47 (2019)
28. Rozario, S.D., Venkatraman, S., Abbas, A.: Challenges in recruitment and selection process: an empirical study. Challenges **10**, 35 (2019)
29. Kotorov, I., Krasylnykova, Y., Zhdanov, P., Mazzara, M.: Finding the right understanding: Twenty-first century university, globalization and internationalization. In: Bruel, J.M., Capozucca, A., Mazzara, M., Meyer, B., Naumchev, A., Sadovykh, A. eds., Frontiers in Software Engineering Education. Lecture Notes in Computer Science, (Cham), pp. 341–353. Springer International Publishing (2020)

30. March, C., Schieferdecker, I.: Technological sovereignty as ability, not autarky, no. 3872378 (2021)
31. Blaschke, P., Demel, J., Kotorov, I.: Innovation performance of Czech and Finnish manufacturing enterprises and their position in the EU. ACC J. **27**, 7–21 (2021)
32. Kappelman, L., Jones, M.C., Johnson, V., McLean, E.R., Boonme, K.: Skills for success at different stages of an it professional"s career. Commun. ACM **59**, 64–70 (2016)
33. Kadoić, N., Ređep, N.B., Divjak, B.: A new method for strategic decision-making in higher education. Central European J. Oper. Res. **26**, 611–628 (2018)
34. Killen, C., Geraldi, J., Kock, A.: The role of decision makers use of visualizations in project portfolio decision making. Int. J. Project Manag. **38**, 267–277 (2020)
35. Kotorov, I., Krasylnykova, Y., Demel, J., Blaschke, P.: The effect of the covid-19 pandemic on economic growth and r&d spending in the eu countries. In: Proceedings of the 16th International Conference, Proceedings of the 16th International Conference, (Liberec, Czech Republic), pp. 68–76, Technical University of Liberec (2023). https://doi.org/10.15240/tul/009/lef-2023-08
36. Agarwal, R., Gaule, P.: What drives innovation? Lessons from covid-19 r&d. J. Health Econ. **82**, 102591 (2022)
37. Alam, G.M., Asimiran, S.: Online technology: sustainable higher education or diploma disease for emerging society during emergency-comparison between pre and during covid-19. Technol. Forecast. Soc. Change **172**, 121034 (2021)
38. Lašáková, A., Bajzíková, Dedze, I.: Barriers and drivers of innovation in higher education: case study-based evidence across ten European Universities. Int. J. Edu. Dev. **55**, 69–79 (2017)
39. Jaleha, A., Machuki, V.: Strategic leadership and organizational performance: a critical review of literature. European Sci. J. ESJ **14** (2018)
40. Romanov, A., Salimzhanov, I., Imam, M., Askarbekuly, N., Mazzara, M., Succi, G., Zhdanov, P., Bobrov, E.: Applying ai in education creating a grading prediction system and digitalizing student profiles. In: 2022 International Conference on Frontiers of Communications, Information System and Data Science (CISDS), pp. 84–93. IEEE (2022)
41. Askarbekuly, N., Solovyov, A., Lukyanchikova, E., Pimenov, D., Mazzara, M.: Building an educational product: Constructive alignment and requirements engineering. In: Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Human Factors in Software and Systems Engineering, Artificial Intelligence and Social Computing, and Energy, 25–29 July 2021, USA, pp. 358–365. Springer (2021)
42. Kloos, C.D., Alario-Hoyos, C., Morales, M., Rocael, Jerez, H.R., Pérez-Sanagustín, M., Kotorov, I., Fernández, S.A.R. Oliva-Córdova, L.M., Solarte, M., Jaramillo, D., Teixeira, A.M., López, A.H.G.: Prof-xxi: teaching and learning centers to support the 21st century professor. In: 2021 World Engineering Education Forum/Global Engineering Deans Council (WEEF/GEDC), pp. 447–454 (2021)
43. Funke, J., Fischer, A., Holt, D.V.: Competencies for complexity: problem solving in the twenty-first century. Educational Assessment in an Information Age, pp. 41–53. Springer International Publishing, Cham (2018)
44. Hanemann, U.: Lifelong literacy: some trends and issues in conceptualising and operationalising literacy from a lifelong learning perspective. Int. Rev. Edu. **61**, 295–326 (2015)
45. Khan, A., Zaib, S., Khan, F., Khan, J., Khan, J., Lee, Y.: Identification and Prioritization of Critical Cyber Security Challenges and Practices for Software Vendor Organizations in Software Development: An AHP-Based Systematic Approach (2022)
46. Blaschke, P., Demel, J., Kotorov, I.: Innovation Performance of Small, Medium-Sized, and Large Enterprises in Czechia and Finland. Technical University of Liberec, Liberec, Czech Republic (2021)
47. Feitosa, J., Hagenbuch, S., Patel, B., Davis, A.: Performing in diverse settings: A diversity, equity, and inclusion approach to culture. International Journal of Cross Cultural Management **22**, 147059582211367 (2022)

48. Driskell, J.E., Salas, E., Driskell, T.: Foundations of teamwork and collaboration. Am. Psychol. **73**(4), 334–348 (2018)
49. Pérez-Sanagustín, M., Kotorov, I., Teixeira, A., Mansilla, F., Broisin, J., Alario-Hoyos, C., Jerez., Teixeira Pinto, M.D.C., García, B., Delgado Kloos, C., Morales, M., Solarte, M., Oliva-Córdova, L.M., Gonzalez Lopez, A.H.: A competency framework for teaching and learning innovation centers for the 21st century: anticipating the post-covid-19 age. Electronics **11**, 413 (2022)
50. Mazzara, M., Zhdanov, P., Bahrami, M.R., Aslam, H., Kotorov, I., Imam, M., Salem, H., Brown, J.A., Pletnev, R.: Education after COVID-19. Advances in Sustainability Science and Technology, pp 193–207. Springer, Singapore (2022)
51. Wojnicka-Sycz, E., Piróg, K., Tutaj, J., Walentynowicz, P., Sycz, P., TenBrink, C.: From adjustment to structural changes—innovation activity of enterprises in the time of covid-19 pandemic. Innov: European J Soc Sci Res **0**(0), 1–26 (2022)

Part VII
Artificial Intelligence, Machine Learning and Deep Learning in Intelligent Decision Technologies

Chapter 35
Skeleton-Based Action Recognition for an Automated Test of Embodied Cognition

Sayda Elmi and Morris Bell

Abstract Cognitive assessment is a critical process aimed at evaluating an individual's cognitive abilities, including memory, attention, problem-solving, and executive functioning. In the context of mental health, cognitive assessment plays a pivotal role in identifying and understanding various neurological disorders such as Alzheimer's, Schizophrenia, and ADHD. One key aspect of cognitive assessment involves measuring executive functioning, which encompasses higher order cognitive processes responsible for planning, organizing, and regulating behavior. To enhance the efficacy of cognitive assessment in real-world applications, our proposed approach, Mind-In-Action (MIA), focuses on skeleton-based action recognition. MIA integrates a sophisticated pose estimator to extract crucial information from human skeletons, enabling automatic measurement of executive functioning through innovative distance and elbow angle calculations. This methodology introduces three distinct score functions—accuracy score, rhythm score, and functioning score—to comprehensively assess and quantify executive functioning in individuals. Through rigorous evaluations on diverse datasets, our MIA model demonstrates significant advancements over existing methods, showcasing its potential as a valuable tool in the field of cognitive assessment.

35.1 Introduction

In the field of video analysis, human action recognition has become increasingly pivotal, garnering significant attention in recent years. Notably, pose estimation studies, as exemplified by [2], have demonstrated promising progress in this domain. To comprehensively capture the spatial configurations and temporal dynamics inherent in human actions, the use of skeleton-based representations has gained prominence.

S. Elmi (✉)
University of New Haven, New Haven, CT, USA
e-mail: saida.elmi@yale.edu

S. Elmi · M. Bell
School of Medicine, Yale University, New Haven, CT, USA
e-mail: morris.bell@yale.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_35

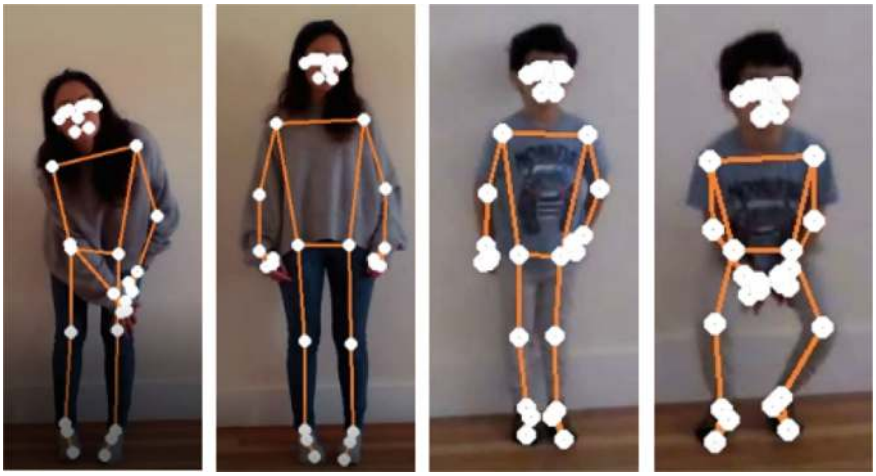


Fig. 35.1 Estimated 2D poses for adults and kids

Human skeletons serve as a concise data format, facilitating the extraction of dynamic features from body movements, as proposed by [8]. In practical applications, human skeletons within videos are typically represented as sequences of joint coordinate lists. While numerous action recognition techniques, as highlighted by [11, 12], showcase impressive performance on public benchmarks where body joint coordinates are obtained through pose estimators. The specific real-world application that we are focusing on in this paper is cognitive assessment in adults and children using cognitively demanding physical tasks. In addition, in the existing state of the art, only the pose information is extracted and skeleton sequences capture only action information while in a real-world application such as cognitive assessment, we need to measure executive functioning in addition to the pose estimation (Fig. 35.1).

Inattention, laziness, hyperactivity, impulsive behavior, lack of motivation, and forgetfulness collectively constitute a spectrum of cognitive impairments prevalent in both children and adults. These manifestations, when observed, often signal underlying challenges in cognitive functioning. Particularly noteworthy is the intricate nature of Attention-Deficit/Hyperactivity Disorder (ADHD), as identified by [13]. Understanding and addressing these diverse cognitive impairments are crucial steps toward effective interventions and support for individuals facing such challenges. Embodied cognition is a well-established construct [15], recognizing that mental functioning involves brain and body working together and cognition develops along with and by way of physical movement. Better measures of cognition that closely relate to individual functioning and predict disability are sorely needed to provide proper remedial intervention at the appropriate time [10]. As measures of embodied cognition, two cognitive games can be used: (i) the Cross-Your-Body (CYB) game can provide sufficient psychometric observations and can be used as a measure of

behavioral self-regulation. As shown in [10], the game is significantly related to cognitive flexibility, inhibitory control, and working memory. The game has four trials with up to four paired behavioral rules: "touch your ears", "touch your shoulders", "touch your hips", and "touch your knees". Subjects first respond naturally, and then are instructed to switch rules by responding in the "opposite" way (e.g., touch their knees when told to touch their ears) and (ii) the Traffic-Lights Game (TLG), which is one of the core tasks with higher cognitive demand, is an attention and response inhibition task. It is similar to computerized continuous performance tests that assess sustained attention and response inhibition but is more complex and requires rhythmic upper body movement in response to commands. The participant is asked to pass a juggling ball (or any other object) from one hand to the other in rhythm to the words "Green Light", to move the ball up and down to the words "Yellow Light" and to not pass the ball when the participant hears "Red Light". The task is subsequently repeated at a faster pace in different trials. The participant is then presented with the same task but using a sequence of pictures of green, red, and yellow traffic lights as visual cues, rather than the spoken cues, thus allowing for comparison between sensory modalities in audio and visual trials. The Traffic-Lights-Game can provide sufficient psychometric observations and can be used as a measure of behavioral self-regulation. The game has four trials and the subjects are expected to perform the sequential movement for every count/beat provided by the therapist. Then, they are evaluated by three scores: (i) Action Score helps to evaluate the working memory and represents the total number of correct actions; (ii) Rhythm Score helps to evaluate the coordination and self-regulation and represents the total number of correct rhythms, accurately keeping the beat. Starting late and rushing to catch the beat is not correct and (iii) Functioning Score helps to evaluate the executive functioning and represents the total number of correct actions in rhythm. Monitoring such a task and scoring it manually is tiresome and requires constant attention from the therapists.

In this paper, inspired by the success of the computer vision tools for action recognition and human skeleton extraction, we investigate how to apply a deep learning model to monitor cognitive abilities, which helps with identifying cognitive impairments. We propose an automated intelligent system called Deep-Cogn, to monitor and assess cognitive behavior through physical tasks which are part of assessment and training for people with Executive Function Disorder. Based on a deep learning model, Deep-Cogn aims to (i) provide an automated infrastructure for performing and evaluating the cognitive assessments. Usually, these assessments are performed by psychologists who manually monitor and score patients which is tiresome and time-consuming; (ii) analyze and predict the action performed by the subjects; and (iii) deliver meaningful information to cognitive experts by providing the subject performance measures like action, rhythm, and functioning scores.

35.2 Problem Formulation

In this section, we first introduce the key data structures used in this paper and formally define our problem. Following the convention, we use capital letters (e.g., X) to represent both matrices and graphs, and use squiggle capital letters (e.g., $\mathcal{X}$) to denote sets. Human action recognition requires to temporally segment all frames of a given video. We first extract a set of temporal regions of interest $\mathcal{X}$ where $\mathcal{X} = X_{i,i\in\{1\ldots K\}}$. Then, we predict the action in each temporal region of interest $X_{i,i\in\{1\ldots K\}}$ and calculate a set of cognitive scores $\mathcal{S}$ where $\mathcal{S} = \{\varphi_A, \varphi_R, \varphi_F\}$ and $\varphi_A, \varphi_R, \varphi_F$ are the action, rhythm, and functioning scores, respectively. The task can be formulated as follows:

Given K temporal regions of interest X, each $X_{i,i\in K}$ is a sequence of D-dimensional features where $X_{i,i\in K} = (x_1, \ldots, x_{|X_i|})$ and $x_{j,j\in\{1..|X_i|\}} \in \mathbb{R}^D$, the task is to infer the sequence of frame-wise action labels $Y_{i,i\in K} = (y_1, \ldots, y_{|Y_i|})$ and there are C action classes $\mathcal{C}$ where $\mathcal{C} = \{1, \ldots, C\}$ and $y_{j,j\in\{1\ldots|Y_i|\}} \in \mathcal{C}$. A set of cognitive scores $\mathcal{S}$ is calculated for every Y_i for K temporal regions of interest.

35.3 Model Architecture

As shown in Fig. 35.2, inspired by [3], the architecture of MIA comprises four major modules which are

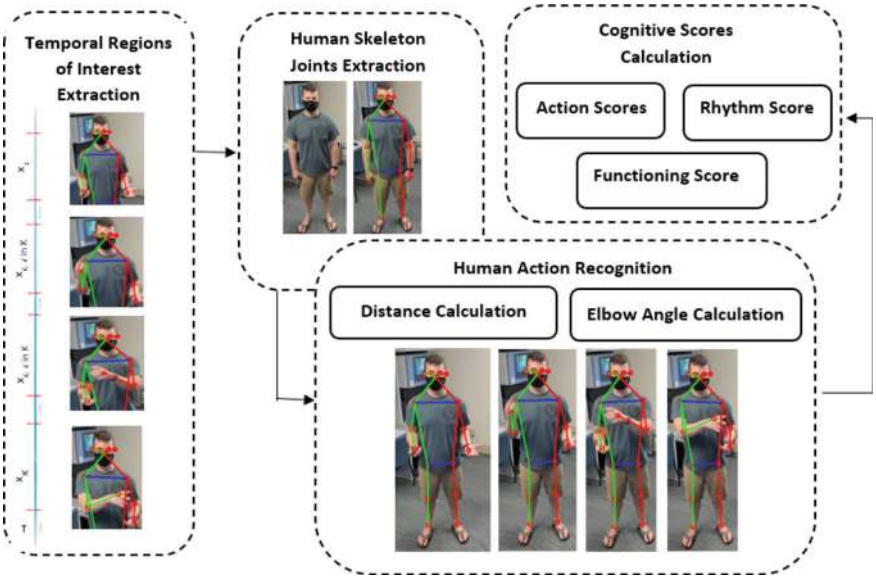


Fig. 35.2 Overview of the proposed architecture

Fig. 35.3 Different classes that occur in the Traffic-Lights task for its four trials

- Temporal regions of interest extraction: Given a trial video, a temporal region of interest (TRI) is defined as the trial segment where the subject is told to do an action. TRI is denoted by $X_{i,i\in K}$ where K is the total number of TRIs in the video.
- Human skeleton joints extraction: For the joint localization problem on RGB data, we explored various existing state-of-the-art methods and decided on using the deep learning architecture proposed in [12]. While some of the other pre-trained models work well with upper body pose [6] and useful for certain other applications, our problem requires an efficient and accurate pose estimator for full-body human joint. In particular, we build upon the GCN model [12] providing highly accurate results regarding the relative position of human body joints. We choose to use the pose estimator model stated above as it is the current state-of-the-art pose estimator for multi-person pose estimation.
- Human action recognition: Based on the Traffic-Lights task, there are a total of three classes, i.e., $|\mathcal{C}| = 3$, as represented in Fig. 35.3 where **class 1**: hands are in the initial position at the same level, when the subject is told "Red-Light", **class 2**: one hand goes up and down when the subject is told "Yellow-Light", **class 3**: one hand is up and ready to pass the ball (or other object) to the other hand.
- Cognitive scoring module: The scoring protocol was created by the psychologist experts and specifies three cognitive scores $\mathcal{S} = \{\varphi_A, \varphi_R, \varphi_F\}$ where $\varphi_A, \varphi_R, \varphi_F$ represent action, rhythm, and function scores, respectively, and they are defined as the amount of times that the subject performs the correct action and reflects how the subject follows the instructions in rhythm.

35.4 Methodology

Temporal Regions of Interest Extraction: The tasks are specifically designed to fit in the theory of embodied cognition, where cognition can be influenced through physical activities. A screen is used to display a music video where the host instructs the subjects to perform the task following the instructions. There are four trials overall and each trial has varying trial segments.

Definition 1 (*Temporal Regions of Interest*) Given a trial video, a temporal region of interest (TRI) is defined as the trial segment where the subject is told to do an action. TRI is denoted by $X_{i,i\in K}$ where K is the total number of TRIs in the video.

While action segmentation requires to temporally segment all frames of a given video [5], we extract K temporal regions of interest X to predict the action in each $X_{i,i\in K}$ as predicting the actions in all frames is more expensive and inefficient. The extraction of the TRIs can be done with the speech recognition to segment the video based on the played instructions. In this paper, the videos are segmented based on the time intervals when the subject is told to perform the action based on the announced instructions as shown in Fig. 35.2.

Human Skeleton Joints Extraction: Currently, skeleton-based representations have been very popular for human action recognition, as human skeletons provide a compact data form to depict dynamic changes in human body movements [11]. Skeleton data is a series of 3D coordinates of multiple skeleton joints, which can be estimated from 2D images [1]. For the joint localization problem on RGB data, we explored various existing state-of-the-art methods and decided on using the deep learning architecture proposed in [12]. While some of the other pre-trained models work well with upper body pose [6] and useful for certain other applications, our problem requires an efficient and accurate pose estimator for full-body human joint. In particular, we build upon the GCN model [12] providing highly accurate results regarding the relative position of human body joints. We choose to use the pose estimator model stated above as it is the current state-of-the-art pose estimator for multi-person pose estimation and also provides competitive performance on single-person pose estimation as demonstrated by their results on popular datasets for these problems. The joints extraction component predicts the location of all human skeleton key points as shown in Fig. 35.2. The tracker predicts the presence of the person on the current frame and the body joint coordinates.

Action Recognition: Human skeletons in a video are mainly represented as a sequence of joint coordinate lists. For the action recognition, we first monitor and analyze hands' positions.

Based on the Traffic-Lights task, there are a total of three classes, i.e., $|\mathcal{C}| = 3$, as represented in Fig. 35.3 where class 1: hands are in the initial position at the same level, when the subject is told "Red-Light", class 2: one hand goes up and down when the subject is told "Yellow-Light", class 3: one hand is up and ready to pass the ball (or other object) to the other hand.

Then, we evaluate if the subject's action complies with the expected motion employing two main measures: (i) **Distance calculation**: The output of our pose tracker is pixel locations for the human body joints. We use these joint locations to measure the distance between hands. (ii) **Elbow angle calculation**: we calculate $\hat{r}$ and $\hat{l}$ as the elbow angles of right hand and left hand, respectively.

A given temporal region of interest X is a sequence of D-dimensional features, called frames, where $X_{i,i\in K} = (x_1, \ldots, x_{|X_i|})$ and $x_{j,j\in\{1\ldots|X_i|\}} \in \mathbb{R}^D$. Action recognition for each temporal region of interest X requires to temporally segment all frames of a given $X_{i,i\in K}$, i.e., predicting the action in each frame $x_{j,j\in\{1\ldots|X_i|\}}$. The probability that a given frame $x_{j,j\in\{1\ldots|X_i|\}}$ belongs to the gesture classes $\mathcal{C}$ is given as follows:

$$P(x_j \in \mathcal{C}) = \begin{cases} 1, & D \leq \alpha \text{ and } A \leq \beta \\ 0 \end{cases}, \tag{35.1}$$

where α and β are the distance and angle thresholds. D is the Euclidean distance between the hand position coordinates. A is the elbow angles where $A = \{\hat{r}, \hat{l}\}$. The gesture classes obtained from the action recognition module are incorporated in the calculations for the cognitive scores in longer sequences of steps performed. The details about calculations of these scores for the steps and sequences that we record are specified in the next sections.

Cognitive Scoring Functions: The scoring protocol created by the psychologist experts specifies three cognitive scores $\mathcal{S} = \{\varphi_A, \varphi_R, \varphi_F\}$ where $\varphi_A, \varphi_R, \varphi_F$ represent action, rhythm, and function scores, respectively, and they are defined as follows: Let Y be the predicted actions for the Traffic-Lights task in a given video.

Definition 2 (*Action Score*) Depends on the amount of times that the subject performs the correct action and defined as

$$\varphi_A(Y) = \sum_{i=1}^{K} y_i \in \mathcal{C}, \tag{35.2}$$

where $\mathcal{C}$ is the set of classes and K is the number of predicted classes in K regions of interest.

We also propose two other cognitive scores, called Rhythm Score and Function Score as defined in [3].

Definition 3 (*Rhythm Score*). Depends on the amount of times that the subject responds within one second after receiving the instruction and defined as

$$\varphi_F(Y) = \sum_{i=1}^{K} y_i \in X_i, \tag{35.3}$$

where X is the set of different regions of interest.

Definition 4 (*Function Score*). Represents the total number of actions that the subject performs the correct actions in the rhythm and defined as

$$\varphi_R(Y) = \sum_{k=1}^{C} \sum_{i=1}^{K} y \in \{X_i \cap C_k\}, \tag{35.4}$$

where C is the number of classes, i.e., $C = |\mathcal{C}|$ and K is the number of temporal regions of interest, i.e., $K = |X|$.

35.5 Experimental Evaluation

In our experiment, we collected data from a broad range of people who are healthy or clinically well characterized and when possible have had conventional cognitive assessments. The adult data collection covers the life-span from 11 year to 90 years. We collected RGB data from 35 participants that are recruited to follow the instructions provided by the music video and perform the task sequences for 4 trials with a total of around 150 videos. Traffic-Lights task was performed twice for a sub-sample, approximately 2 weeks apart. Motion capture data were collected and then converted into cognitive scores.

Our MIA model aims to predict three cognitive scores, i.e., φ_A, φ_R, and φ_F. We measure our method by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for each of our predicted cognitive scores as follows: RMSE $= \sqrt{\frac{1}{n}\Sigma_{t=1}^{n}\Big(s - \varphi(Y)_t\Big)^2}$ and MAE $= \frac{1}{n}\Sigma_{t=1}^{n}|s - \varphi(Y)_t|$ where $\varphi(Y)$ and s are the predicted score value and real score value, respectively; n is the number of all predicted score values. For the first three trials, the subjects are required to follow the audio instructions, but for the last trial, the challenge becomes cognitively demanding as they are told to follow visual instructions showing green, yellow, and red traffic lights. In the following, we evaluate the performance of our MIA model in terms of accuracy of the predicted cognitive scores. The results on four trials were used to evaluate the accuracy, reported in Table 35.1. Four deep learning-based methods are used to evaluate our proposed model MIA: (i) MSTCN [4] introduces an auxiliary self-supervised task to find correct and in-correct temporal relations in videos using smoothing loss to avoid over-segmentation errors, (ii) DTGRM [14] uses Graph Convolution Networks (GCN) and to model temporal relations in videos. It has the ability for efficient temporal reasoning, (iii) MSTCN++ [9] is an improvement over MSTCN where the system generates frame level predictions using a dual dilated layer that combines small and large receptive field and (iv) ASRF [7] alleviates over-segmentation errors by detecting action boundaries. Table 35.1 shows the comparative performances for MSTCN, DTGRM, MSTCN++, ASRF, and MIA for different predicted cognitive scores. On Adult datasets, MSTCN++ shows a better performance against DTGRM and ASRF while MIA has the best performance comparing to other models, being

Table 35.1 Evaluation of MIA model in terms of RMSE and MAE

Method	Adult dataset					
	Action score		Rhythm score		Function score	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
DTGRM	1.652	0.791	1.521	1.192	0.101	0.498
MSTCN++	1.532	0.738	1.422	1.132	0.925	0.412
ASRF	1.743	0.818	1.635	1.254	1.132	0.52
MSTCN	2.051	1.215	1.982	1.458	1.532	0.891
MIA	**1.312**	**0.635**	**1.198**	**0.912**	**0.707**	**0.250**

immune to contextual nuisances, such as background variation and lighting changes. Figure 35.4 shows the results on kid datasets. We evaluate the accuracy (both RMSE and MAE are reported) of the different cognitive scores, i.e., action, rhythm, and function scores, for the four trials. Results show that prediction for function score is highly related to action score prediction. That is because when the actions are misidentified, both scores are impacted. Although the actions performed by children might be imprecise, results on Kid datasets show similar variations over cognitive

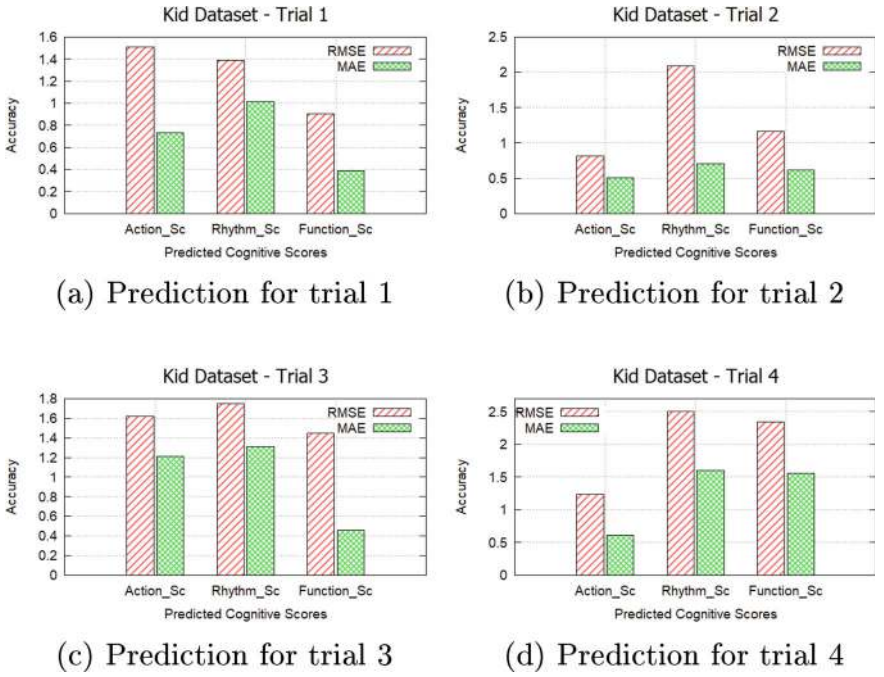


Fig. 35.4 Kid dataset: prediction for different trials

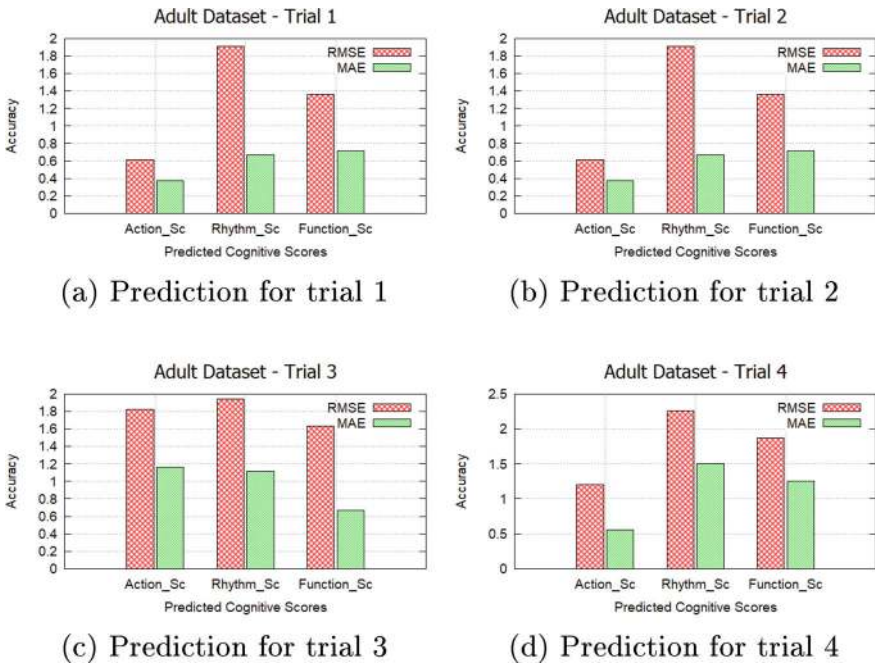


Fig. 35.5 Adult dataset: prediction for different trials

scores as reported in Fig. 35.4. That means that the two measures, distance and elbow angle calculation, both have an impact to deal with the fine-grained motions (Fig. 35.5).

35.6 Conclusion

This paper proposes a deep learning-based action recognition method for evaluating and monitoring cognitive abilities of human subjects. We deploy a deep learning architecture to analyze human activity and provide informative measures to the experts regarding the performance of the subject, i.e., cognitive scores.

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: Openpose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
2. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. CVPR (2022)

3. Elmi, S., Bell, M., Tan, K.-L.: Skeleton-based human action recognition for cognitive behavior assessment. In: ICTAI, Deep-cogn (2022)
4. Farha, Y.A., Gall, J.: MS-TCN: multi-stage temporal convolutional network for action segmentation. In: IEEE CVPR, pp. 3575–3584 (2019)
5. Fayyaz, M., Gall, J.: SCT: set constrained temporal transformer for set supervised action segmentation. In: CVPR, pp. 498–507 (2020)
6. Gattupalli, S., Ghaderi, A., Athitsos, V.: Evaluation of deep learning based pose estimation for sign language recognition. In: PETRA (2016)
7. Ishikawa, Y., Kasai, S., Aoki, Y., Kataoka, H.: Alleviating over-segmentation errors by detecting action boundaries. In: WACV (2021)
8. Johansson, G.: Visual perception of biological motion and a model for its analysis. Percept. Psychophys. 14(2) (1973)
9. Li, S., Farha, Y.A., Liu, Y., Cheng, M.-M., Gall, J.: MS-TCN++: multi-stage temporal convolutional network for action segmentation
10. McClelland, M.M., Cameron, C.E., Duncan, R., Bowles, R.P., Acock, A.C., Miao, A., Pratt, M.E.: Predictors of early growth in academic achievement: the head-toes-knees-shoulders task. Front. Psychol. **5**(2) (2014)
11. Song, Y.-F., Zhang, Z., Shan, C., Wang, L.: Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition. In: MM, pp. 1625–1633. ACM (2020)
12. Song, Y.-F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Patterns Pattern Anal. Mach. Intell. (2021)
13. Tucha, O.: The history of attention deficit hyperactivity disorder. ADHD Atten. Deficit Hyperact. Disord. **2**(4) (1999)
14. Wang, D., Hu, D., Li, X., Dou, D.: Temporal relational modeling with self-supervision for action segmentation. CoRR. arXiv:2012.07508 (2020)
15. Wilson, A.D., Golonka, S.: Embodied cognition is not what you think it is. Front. Psychol. **12**(2) (2013)

Chapter 36
Advancing Gender Equality in Media: Tackling Stereotypes and Biases with AI

Zhan Liu, Anne Darbellay, Nicole Glassey Balet, and Valérie Vuille

Abstract This study presents an innovative approach to evaluating media representations of gender-based violence by integrating Natural Language Processing (NLP) techniques with the advanced capabilities of GPT-4, an Artificial Intelligence (AI)-based large language model. We developed a set of 27 expert-defined criteria to analyze a corpus of news articles, initially utilizing NLP methods for foundational text analysis. For more complex criteria, we employed GPT-4 and further enhanced its precision with fine-tuning. Our results indicate a significant increase in accuracy, achieving an overall 76% accuracy rate in content evaluation, which is 9% points higher than using NLP alone. This research introduces a novel media content analysis framework and paves the way for future enhancements in automated journalism assessment and ethical reporting.

36.1 Introduction

Gender-based violence, a pervasive global crisis deeply ingrained in gender inequality and unjust power dynamics, profoundly affects millions, especially women, girls, and the LGBTIQ+ community. According to DécadréE [9], this violence includes any act committed against a person's will, reflecting society's binary gender roles and unequal power relations. It encompasses threats and coercion, and can manifest as physical, emotional, psychosocial, or sexual violence, including deprivation of

Z. Liu (✉) · A. Darbellay · N. G. Balet
Media Innovation Lab, University of Applied Sciences and Arts Western Switzerland, HES-SO Valais-Wallis, Delémont, Switzerland
e-mail: zhan.liu@hevs.ch

A. Darbellay
e-mail: anne.darbellay@hevs.ch

N. G. Balet
e-mail: nicole.glassey@hevs.ch

V. Vuille
DécadréE, Geneva, Switzerland
e-mail: valerie.vuille@decadree.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_36

resources or access to services. Notably, women, minors, and those whose gender identity or sexual and affective orientation diverge from societal norms are predominantly affected. The World Health Organization's 2021 [20] data indicates that roughly one in three women have experienced such violence, either from an intimate partner or another individual. The multifaceted nature of this violence, ranging from psychological abuse and physical harm to femicide, underscores the importance of media portrayal. Media representations can either perpetuate harmful stereotypes or contribute to a more nuanced and empathetic understanding. Thus, comprehending the complex dynamics of gender-based violence is vital as we explore how AI-driven analysis and intervention in media reporting can sensitively challenge biases and influence public perception and policy.

The media has been instrumental in bringing gender-based violence into public discourse, highlighting both high-profile cases and the daily struggles of survivors. However, this coverage is not without its flaws, as it often falls into the trap of perpetuating harmful stereotypes and biases. Women stepping forward with sexual assault allegations frequently face victim-blaming and skepticism, while male perpetrators sometimes receive sympathetic portrayals, especially if they are prominent figures. Additionally, the media's tendency to focus on sensationalist and extreme incidents can overshadow the broader social and cultural contributors to gender-based violence. This skewed portrayal underscores the urgency to rethink how gender-based violence is reported and to foster approaches that ensure balanced and equitable coverage.

In response to these issues, the Council of Europe [7] has urged member states to establish legal frameworks that uphold human dignity and prevent gender-based discrimination and violence. Media organizations are also encouraged to implement self-regulatory systems and ethical guidelines to support gender equality and eliminate discriminatory content. However, despite these initiatives, inconsistencies and biases in media language and imagery persist, undermining efforts to combat gender stereotypes and sexist violence. AI holds the potential for transformative impact in this area. By integrating tools like machine learning, natural language processing, and sentiment analysis into media practices, AI can critically assess and adjust how gender-based violence is reported. These AI-driven tools can identify patterns of bias, assist in creating more gender-neutral content, and provide insights into diverse gender representations, promoting a more inclusive media landscape.

In this study, we developed an automated, artificial intelligence-based content evaluation system, designed to streamline and enhance the analysis of gender-based violence in media reports. Collaborating with large language models from OpenAI,[1] our system utilizes sophisticated natural language processing techniques to automatically analyze and evaluate media coverage from various sources. The integration of comprehensive processes—including data collection, preprocessing, feature engineering, model design, and the establishment of specific evaluation criteria—enables our system to function as a highly efficient decision-making aid. Through rigorous

[1] https://openai.com/.

evaluating across 27 different gender equality evaluation criteria, our model demonstrated an average accuracy of 76%, with 8 criteria having peak accuracy above 90%. This research has garnered recognition from field experts for its significant contributions to enhancing work efficiency and promoting gender equality in news content, demonstrating the substantial potential of AI in transforming media practices.

The paper is structured to succinctly present our research and findings. Section 36.2 provides a literature review, framing our work within existing research. Section 36.3 details the dataset description, including data collection and preparation. Section 36.4 outlines our methodology, evaluation criteria, and implementation process, while Sect. 36.5 presents the experimental setup and results, demonstrating our system's effectiveness. The conclusion summarizes our contributions and proposes directions for future research.

36.2 Related Work

Extensive research has scrutinized the media's portrayal of gender-based violence, revealing a dual narrative. On one hand, studies [4–6] criticize media coverage for often sensationalizing extreme cases, perpetuating harmful stereotypes, engaging in victim-blaming, and sympathetically portraying perpetrators, thereby distorting the realities of gender-based violence. On the other hand, the media is acknowledged as a pivotal platform for raising awareness and advocating for societal change, particularly when adopting ethical standards and self-regulatory codes [10, 11, 14]. DecadréE [9] underscores the complexity of this issue, attributing problematic media representations to a blend of individual, structural, and systemic factors, and emphasizes the need for a nuanced approach in addressing these narratives to foster fairer and more consistent reporting on gender-based violence.

The use of innovative technologies, specifically AI and NLP, in media content analysis is a rapidly growing field of study. Recent literature [3, 8, 13] indicated that machine learning and AI can be powerful tools for analyzing large volumes of media content, identifying patterns, and evaluating the quality of reporting. Such techniques can automatically categorize media content, detect biases, and evaluate how well the media adheres to ethical standards and practices, thus providing a scalable solution for media content evaluation.

In the context of gender-based violence, AI and NLP could potentially be used to identify problematic representations and language use in media coverage, thereby aiding in the promotion of gender equality and combating sexist violence. Existing research has shown that AI can be trained to understand complex language patterns and detect subtle biases that may be challenging for humans to identify. This could be particularly beneficial for understanding how gender-based violence is framed and discussed in the media. However, most of the current research [1, 18] focused on the classification of news content and does not address the evaluation of gender-based violence content from different media.

Moreover, some studies highlight the impact of AI-based solutions on journalism and content creation [2, 16, 19]. These solutions could provide insights into journalistic practices and raise awareness about the importance of fair and equitable reporting. This has implications for improving media practices, promoting gender equality, and ultimately contributing to the fight against gender-based violence.

Nevertheless, as with any technological solution, there are challenges and ethical considerations associated with using AI and NLP for media content analysis. Concerns around algorithmic bias, transparency, and the implications for freedom of speech and journalistic autonomy are prevalent in the literature. Therefore, we need to be careful about the design and implementation of such systems to ensure they are used responsibly and ethically.

36.3 Dataset Description

This study utilizes a specifically curated dataset from the Swissdox database,[2] a comprehensive collection of about 24 million media articles from a wide range of Swiss media sources covering a substantial temporal range. To construct a dataset specifically relevant to our study of gender-based violence, we employed a keyword-based filtering approach. This process involved selecting articles from Swiss French-language media sources using keywords indicative of various facets of gender-based violence, including "violence against women", "sexual harassment", "street harassment", "rape", "marital drama", "family drama", "domestic violence", "mutilation", "femicide", and "sexual coercion". We designed these selection criteria to ensure the retrieval of articles directly pertinent to our research theme.

Our data compilation yielded a collection of 1,752 news articles from 19 distinct French-speaking media outlets in Switzerland for the year 2022, including prominent publications like "Le Courrier", "Le Temps", "RTS Info", "24 Heures", "Le Nouvelliste", "20 Minutes", "Le Matin", "Swissinfo", and "Blick". Prior to analysis, we performed several preprocessing steps on the dataset: standardizing the format by removing all HTML tags and excluding duplicates and incomplete articles, resulting in a refined subset of 1,046 articles. Each article was categorized with a title, summary, and main body. The titles had an average length of 11.5 words (SD = 2.92), with a range of 4–21 words. The body text of the articles was substantially longer, averaging 508.5 words (SD = 332.67), with the shortest at 15 words and the longest at 3,757 words.

A focused subset of 1,046 articles underwent detailed annotation based on 27 expert-defined criteria, which provided a nuanced framework for our analysis of gender-based violence representation in the media. This rigorous annotation was pivotal for a thorough examination of these portrayals within the Swiss French-language media landscape. It also established a benchmark for evaluating the performance of our AI-based content evaluation system. By measuring the AI system's

[2] https://swissdox.ch/.

outputs against the expert annotations, we could determine the system's accuracy and its capability to emulate expert analysis. This step was crucial in validating the AI system's utility for media content analysis, with a particular emphasis on identifying and assessing the nuanced depictions of gender-based violence.

36.4 Methods and Implementation of Content Evaluation System

In this section, we detail the methodologies and processes used in creating and applying our AI-based system for evaluating news media portrayals of gender-based violence. We outline the step-by-step technical and procedural aspects, from data collection to AI system evaluation. The following subsections will describe our workflow, evaluation criteria, and the systematic implementation of our solution, highlighting the methodological rigor and innovation at the core of our research.

36.4.1 Architecture

The architecture of our content evaluation system is underpinned by a three-tiered workflow (as shown in Fig. 36.1), designed to streamline the process from data acquisition to the synthesis of actionable insights. Each phase is interlinked, forming a cohesive pipeline that underwrites the robustness of our analysis.

Our system's foundation was built upon meticulously acquiring data from the SWISSDOX database, rich in Swiss media articles. We strategically selected sources from French-speaking Swiss news media using keywords related to gender-based violence, ensuring the dataset's relevance. Following selection, we embarked on a rigorous data cleaning process, purifying the dataset to ensure a high quality of information for subsequent analysis.

The second phase of our workflow involved processing and analyzing the prepared data. We began with data extraction and annotation, tagging content based on predefined criteria. Following this, data modeling structured the information for detailed analysis. We employed advanced AI and NLP technologies for a deeper understanding and interpretation of the corpus, integrating external knowledge bases to enhance and contextualize our analysis further.

In the final phase, we defined a comprehensive set of evaluation criteria, against which the media content was assessed. We then moved to the implementation stage, where a prototype of our content evaluation system was developed. This prototype was rigorously tested to ensure its efficacy in automatically evaluating gender-based violence content in news media. To ensure the validity of our system, we engaged

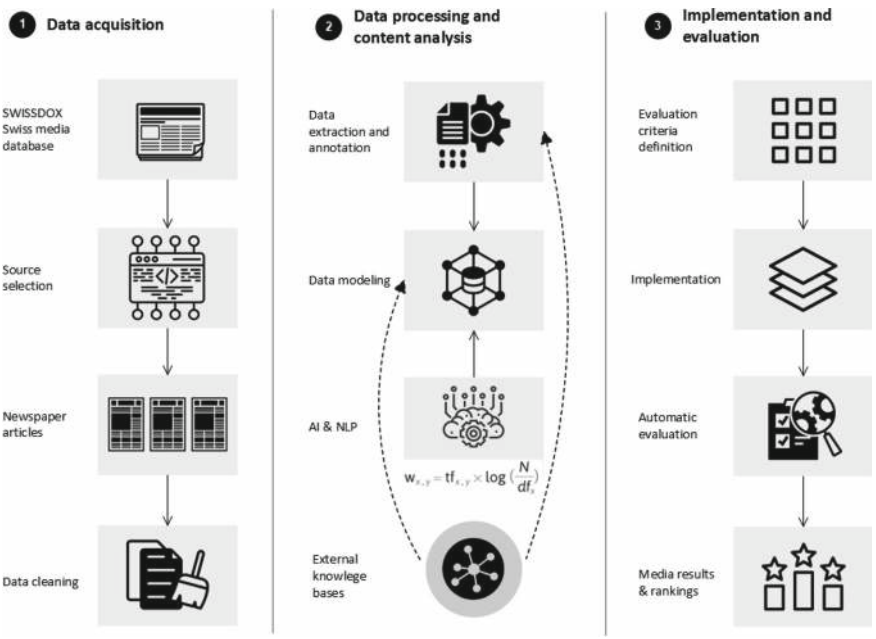


Fig. 36.1 Architecture of news content evaluation system

in an automatic evaluation process, where the system's outputs were compared against benchmarks set by domain experts. Finally, we conducted results analysis and ranking, assessing the performance of various media outlets in their coverage of gender-based violence, and highlighting areas for potential improvement.

36.4.2 Evaluation Criteria

Building upon our foundational methodology, which was informed by 27 expert-defined criteria from DécadréE [9], we embrace a broad spectrum of violence types, including psychological, economic, physical, and sexual violence, as outlined by [12]. This broad approach allows us to cover various manifestations of gender-based violence, such as insults, denigration, and economic deprivation. Nonetheless, despite the inclusion of "symbolic" violence in studies like [17], challenges in attributing specific perpetrators led us to focus our analysis on more directly identifiable forms of violence, adhering to our methodological strengths and the practical scope of our research.

The evaluation began by verifying the presence of fundamental metadata in each article, such as the publication date and the author's name, which are cornerstones of journalistic credibility. We assessed the diversity of media sources, specifically

identifying each article's origin from our list of 19 French-speaking news outlets. This step was critical for understanding the landscape of the reporting and any potential media-specific biases.

We then analyzed the content's relevance, categorizing articles by their direct relation to gender-based violence, and classifying them by length to infer the detail level provided. The type of violence reported was identified, noting especially the mention of femicide, to determine the focus areas of media attention.

The language used within articles underwent meticulous examination. We looked for the presence of certain vocabularies that could either clarify or obfuscate the severity of the incidents. This included terms that could potentially romanticize the violence or attribute it to factors like "passion", which can be misleading and harmful.

The portrayal of individuals involved in reported incidents was another focal point. Descriptions of the victim's and perpetrator's behaviors were evaluated, alongside any mentions of the perpetrator's nationality or mental health, which are often laden with stereotypes. The inclusion of such details was noted for its potential to influence public perception and perpetuate stigmas.

Our criteria also extended to the analysis of articles for underlying themes that are pivotal in understanding gender-based violence. We checked for discussions on power dynamics, control mechanisms, and the escalation of violence. Furthermore, we considered whether articles connected individual incidents to wider societal issues, such as rape culture, and whether they employed statistical data to contextualize the violence within a broader societal framework.

Finally, we considered the articles' resourcefulness to the public, noting the inclusion of information beneficial to victims, like support hotlines or educational websites. The presence of expert interviews was also a key criterion, as it can lend authority and depth to the articles, providing readers with a more informed perspective on the subject matter.

Through the application of these expert-defined criteria, we aimed to provide a rigorous and nuanced assessment of news content, aiming to elevate the standard of reporting on gender-based violence and contribute to a more informed and empathetic public discourse.

36.4.3 Implementation Process

Our content evaluation system represents an intricate fusion of NLP techniques and the nuanced understanding capabilities of AI-based large language models like ChatGPT. Initially, we employed foundational NLP methods—lemmatization, tokenization, POS tagging, and Named Entity Recognition—to structure the data efficiently. To quantify the significance of specific terms within our corpus, we integrated the TF-IDF weighting scheme, calculated using $W_{xy} = TF_{xy} \times \log\left(\frac{N}{df_x}\right)$. This formula, where TF_{xy} represents term frequency and $\log\left(\frac{N}{df_x}\right)$ signifies the inverse document

frequency, helps in moderating a term's weight based on its occurrence, ensuring a balanced emphasis on contextually relevant terms.

Beyond these traditional methods, we addressed more complex evaluation criteria by employing large language models, recognizing their adeptness in processing extensive data and discerning subtle narrative nuances. Specifically, these models excelled in analyzing societal implications of gender-based violence and understanding intricate power dynamics between involved parties, tasks that extend beyond the scope of conventional NLP.

To further refine our system's performance, we fine-tuned the large language model (LLM), a process that significantly enhanced its predictive accuracy. Our fine-tuning process involved adjusting the LLM model's parameters specifically for the context of gender-based violence in media content. This was achieved by initially training the model on a broad dataset of general text to establish a baseline understanding. Subsequently, we introduced a specialized dataset comprised of articles related to gender-based violence, annotated by experts with the nuances and complexities specific to this subject matter. The training process utilized a smaller learning rate to make subtle adjustments, ensuring the model can simulate natural human responses, to become more aligned with the thematic content of our research. We also implemented validation checks to monitor the model's performance and avoid overfitting, thereby maintaining its ability to generalize across different contexts while being adept at recognizing and analyzing the specific patterns of gender-based violence in media reporting.

Integrating this enhanced model with our robust NLP framework allowed us to construct a system of remarkable analytical depth. This synergy between fine-tuned AI models and rigorous NLP techniques enabled us to conduct a granular yet comprehensive examination of news content. Consequently, our system proved exceptionally adept at navigating the complex narrative landscape of gender-based violence, providing insights with unprecedented precision and depth. This advanced analytical capacity ensured a thorough and nuanced exploration of media reporting, setting a new standard for automated content analysis in the context of social and cultural research.

36.5 Experimental Setup and Results

In this section, we describe the environment used to perform our experiments and the accuracy results through our methodologies.

36.5.1 Environment Setup

Our experimental environment leveraged the Python programming language, known for its versatility and robust library ecosystem in data science and machine learning.

Central to our computational work was the IPython framework [15], which facilitated interactive scientific computing and dynamic code execution, crucial for our complex data analysis tasks.

For natural language processing, we employed Spacy,[3] an open-source NLP library, which provided the necessary tools for efficient and in-depth text analysis. The core of our computational analysis hinged on the OpenAI API's GPT-4 large multimodal model,[4] renowned for its superior performance in text generation and understanding. Initially, we performed TF-IDF and classification calculations using the Scikit-learn toolkit,[5] but shifted to GPT-4 due to its enhanced accuracy in handling complex analytical tasks. In addition, we applied the Pandas[6] and Numpy[7] libraries for data processing and numerical analysis. These tools were instrumental in managing, processing, and analyzing our dataset, ensuring a streamlined and effective experimental setup.

36.5.2 Experimental Results

In our evaluation of media content against the 27 defined criteria, we initially utilized various NLP and classification methods to analyze the dataset. Our findings indicated that the results for 12 of these criteria were below our threshold for satisfaction. Consequently, we turned to advanced models, specifically GPT-4 and its fine-tuned counterpart, to reassess these 12 criteria. The comparative analysis revealed a marked improvement in accuracy when employing these AI models.

Figure 36.2 illustrates some examples of the accuracy levels achieved by different methods. The vertical axis quantifies the accuracy, ranging from 0 to 1, while the horizontal axis categorizes the methods into NLP, GPT-4, and GPT-4 with fine-tuning. Notably, the fine-tuned GPT-4 model consistently outperformed the others, underscoring the value of model customization.

When considering the entire set of 27 criteria, we observed a significant variance in accuracy rates, with the highest-performing criterion reaching 96% accuracy and the lowest at 44% by using the model of GPT-4 with fine-tuning. Our analysis through the GPT-4 model highlighted that article length was a determining factor in accuracy, longer articles tended to yield lower accuracy rates. Additionally, the style of writing influenced the results, a simpler narrative structure facilitated higher accuracy levels. Through the integration of GPT-4 models, we achieved an overall accuracy of 76% in our news article analysis, an improvement of nine percentage points over the methods relying solely on NLP. This enhancement demonstrates the effectiveness

[3] https://spacy.io/.

[4] https://openai.com/gpt-4.

[5] https://scikit-learn.org/.

[6] https://pandas.pydata.org/.

[7] https://numpy.org/.

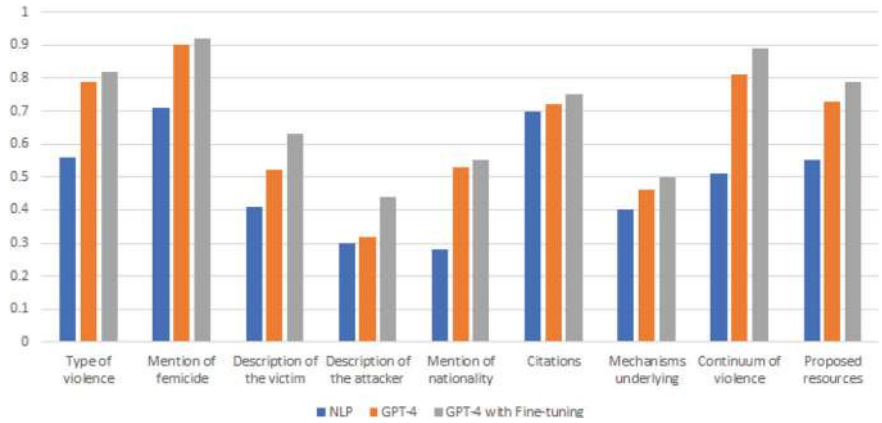


Fig. 36.2 Accuracy of criteria evaluation with different methods

of incorporating AI-based large language models and fine-tuning techniques in the context of complex content analysis tasks.

36.6 Conclusion

In this study, we have demonstrated the efficacy of integrating advanced NLP methods with AI-driven large language models, particularly GPT-4, to evaluate media content related to gender-based violence. Our hybrid approach utilized traditional NLP techniques for initial text analysis, followed by the application of GPT-4 models, fine-tuned to our specific dataset, to handle more complex analytical tasks. The results signify a considerable improvement in accuracy, particularly in analyzing criteria that traditional NLP methods found challenging.

Our contributions through this research are twofold: First, we have provided a comprehensive set of evaluation criteria grounded in expert knowledge, tailored to assess the quality of media reporting on gender-based violence. Second, we have established a methodological framework that combines the best of NLP and AI technologies to create a powerful tool for content analysis. The success of our approach is reflected in the precision of our system, which outperformed standard NLP techniques by a significant margin.

For future research directions, expanding the system's capabilities to handle broader and more diverse datasets could offer insights into global media narratives. Enhancing the fine-tuning process with a variety of large language model architectures could improve accuracy. Additionally, broadening the evaluation criteria to address new themes in media, such as intersectionality within gender-based violence contexts, presents a valuable area for study. Ethical considerations around AI's use in

media analysis and the potential for real-time feedback mechanisms for media outlets also warrant further exploration, aiming for transparency, fairness, and improved reporting standards.

In conclusion, our research has not only advanced the current understanding and methodology of media content analysis but has also laid the groundwork for future innovations that can further enhance the quality and ethical standards of journalism.

Acknowledgements The research presented in this article was supported by a grant from the University of Applied Sciences and Arts of Western Switzerland (HES-SO) focused on the special topic of Gendered Innovation, under grant number 128298. We extend our sincere gratitude to our project partner DécadréE, for the invaluable support of the project.

References

1. Abdulkareem, L.R., Karan, O.: Using ANN to predict gender-based violence in Iraq: how AI and data mining technologies revolutionized social networks to make a safer world. In: 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 298–302. IEEE (2022)
2. Bailer, W., Thallinger, G., Krawarik, V., Schell, K., Ertelthalner, V.: AI for the media industry: application potential and automation levels. In: International Conference on Multimedia Modeling, pp. 109–118. Springer (2022)
3. Bello, H.J., Palomar, N., Gallego, E., Navascués, L.J., Lozano, C.: Machine learning to study the impact of gender-based violence in the news media. arXiv:2012.07490 (2020)
4. Berns, N.: Degendering the problem and gendering the blame: political discourse on women and violence. Gender Soc. **15**(2), 262–281 (2001)
5. Buiten, D.: Silences stifling transformation: misogyny and gender-based violence in the media. Agenda: Empower. Women Gender Equity (71), 114–121
6. Cuklanz, L.: Representations of gendered violence in mainstream media. Quest. Commun. **35**, 307–321 (2019)
7. de Europa, C.: Convention on preventing and combating violence against women and domestic violence (2011)
8. de Lima-Santos, M.F., Ceron, W.: Artificial intelligence in news media: current perceptions and future outlook. Journal. Media **3**(1), 13–26 (2021)
9. Décadrée annual report: Media treatment of gender-based violence. https://decadree.com/wp-content/uploads/2020/09/rapport-2020.pdf
10. Journalism is an essential lever in the fight against gender-based violence. https://www.unicef.org/moldova/en/blog/journalism-essential-lever-fight-against-gender-based-violence. Accessed 11 December 2023
11. Jukic, E.: Research on media reporting on gender based violence against women in Bosnia and Herzegovina. Retrieved on March **25**, 2019 (2016)
12. Kelly, L.: The continuum of sexual violence. In: Women, Violence and Social Control, pp. 46–60. Springer (1987)
13. Liu, Z., Balet, N.G.: Bringing big data into media: a decision-making model for targeting digital news content. In: 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD), pp. 218–223. IEEE (2022)
14. Media: A key to addressing and ending gender-based violence (GBV). https://catalystasconsulting.com/media-a-key-to-addressing-and-ending-gender-based-violence-gbv/. Accessed 11 December 2023

15. Pérez, F., Granger, B.E.: Ipython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007)
16. Pihlajarinne, T., Alen-Savikko, A.: Introduction to artificial intelligence, media and regulation. In: Artificial Intelligence and the Media (2022)
17. Sepulchre, S., Manon, T.: La représentation des violences sexistes et intrafamiliales dans la presse écrite belge francophone. Technical report (2019)
18. Soldevilla, I., Flores, N.: Natural language processing through BERT for identifying gender-based violence messages on social media. In: 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), pp. 204–208. IEEE (2021)
19. Tejedor, S., Vila, P.: Exo journalism: a conceptual approach to a hybrid formula between journalism and artificial intelligence. Journal. Media **2**(4), 830–840 (2021)
20. Violence against women. World health organization. https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Chapter 37
IoT-Enhanced Tomato Leaf Disease Identification Using MLP-Mixer in Agricultural Environments

Besma Rabhi, Habib Dhahri, Imen Jdey, and Omar Alhajlah

Abstract Tomato diseases cause large output losses and represent a serious danger to the world's economy and productivity. Manual evaluation is expensive and difficult. Modern farming methods can be replaced with more effective ones by utilizing artificial intelligence and cutting-edge technology. The objective of this research is to create a Deep Learning (DL) technique that can precisely identify tomato diseases from images of the leaves in an Internet of Things (IoT) environment. Using the PlantVillage Dataset as a benchmark, the proposed MLP-Mixer model was assessed with other well-known transfer learning methods, including VGG16, AlexNet, VGG19, and FNet. A comparison of each model's parameter complexity and performance assessments was part of the analysis. With an area under the characteristic curve of 99.69% and a peak accuracy of 98.34%, the MLP-Mixer method produced impressive results.

37.1 Introduction

Food and Agriculture Organization (FAO) [1] estimates that between 20 and 40% of the world's crop yield is lost to pests and insects. In contrast to creepy crawlies, which cost the world economy about $70 billion year, plant diseases cost about $220 billion annually. Artificial Intelligence (AI) was used to monitor temperature, humidity, climate, satellite data, and weather forecasts in order to increase irrigation efficiency. Within the existing literature, several contributions have been made in the field of tomato disease identification. In [2], the authors introduced a deep learning

B. Rabhi (✉) · I. Jdey
REsearch Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, 3038 Sfax, Tunisia
e-mail: besma.rebhi.2015@ieee.org

H. Dhahri · O. Alhajlah
College of Applied Computer Sciences (CACS), King Saud University, Riyadh 11543, Saudi Arabia
e-mail: hdhahri@ksu.edu.sa

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_37

model that integrates attention modules and residual blocks to distinguish between healthy and diseased tomato leaves. Their study, conducted on the widely recognized PlantVillage dataset, demonstrated superior performance compared to previous deep learning models [3, 4].

Another significant work, presented in [5], proposed the EfficientNet architecture, utilizing both the PlantVillage dataset and an augmented dataset with 55,448 and 61,486 photos, respectively. The study revealed that the B5 and B4 models from the EfficientNet architecture achieved top scores in disease classification. In [6], a straightforward transfer learning-based method for identifying tomato leaf diseases was presented. The authors applied pre-processing techniques to enhance leaf photos by adjusting illumination, using a previously trained MobileNetV2 model for feature extraction and disease classification. The authors in [7] proposed a hybrid deep learning model for classifying tomato diseases, incorporating PCA-whale optimization. Following the extraction of the dataset's most significant features through the whale optimization process, the PCA technique is applied to reduce the data volume. The study applied these findings to the PlantVillage dataset. Subsequently, the proposed model from this work is compared to well-known machine learning techniques. The collected data indicate that the model outperforms others in terms of both accuracy and efficiency.

Smart Farming integrates advanced technologies like cloud computing and the Internet of Things (IoT) to enhance the agricultural management cycle. This integration facilitates the use of robots and artificial intelligence, programmed for tasks such as quality inspections, plant/weed detection, and efficient crop harvesting. A notable focus is on weed management. Additionally, the integration of deep learning techniques with IoT technology stands out as a highly promising method for early identification of leaf diseases and supporting crop improvement.

The accessibility of extensive training data and advancements in technology has facilitated the advancement of CNN research. Simultaneously, the introduction of new architectural concepts and strategies for parameter optimization has expedited research progress, enabling the widespread application of CNNs across various domains [8, 9].

Recently, self-attention operation of Vision Transformer (ViT) [10, 11] and its variations have replaced convolution, demonstrating accuracy levels equivalent to or surpassing CNNs. The MLP-Mixer's architecture, distinctively, adopts an MLP-only approach without convolution or self-attention processes. To facilitate cross-patching communications, it introduces a token-mixing MLP alongside the channel-mixing MLP [12]. While attention mechanisms and transformer models have dominated deep learning, recent observations suggest that attention alone may not provide sufficient power. Attempts to substitute attention for the feed-forward network in a transformer have resulted in notably poor performance, indicating a potential shortfall in attention-based models.

Relying on MLPs instead of inductively biased models like CNNs is feasible due to two main factors: (1) the availability of sufficient data and (2) the utilization of robust optimization, normalization, and Data Augmentation (DA) techniques [8]. Unlike CNNs and transformers, which entail quadratic costs, the MLP-Mixer presents a

linear computational complexity concerning the total number of input patches. A key advantage lies in the fact that, despite its significantly faster processing, the Mixer outperforms competing systems.

In summary, the contributions of the paper are as follows:

- We developed a specialized deep learning method dedicated to accurately identifying tomato diseases from leaf photos within an IoT framework.
- We integrated the proposed MLP-Mixer model into an Internet of Things (IoT) framework, showcasing the potential of advanced technology in improving agricultural disease detection.
- We conducted a rigorous evaluation of the MLP-Mixer model against the benchmark PlantVillage Dataset, establishing a standardized basis for performance assessment, and performed a comprehensive comparative analysis, contrasting the MLP-Mixer model with established transfer learning models (VGG16, AlexNet, VGG19, FNet) to highlight its relative strengths.
- We achieved a balanced model, maintaining a high level of accuracy while minimizing the parameter count, emphasizing its practical utility for farmers in early disease detection.

The subsequent sections of the paper are organized as follows: Sect. 37.2 delineates the proposed methodologies. Section 37.3 contrasts the results derived from various models with the proposed model. Section 37.4 provides a summary of the work and proposes potential avenues for further research.

37.2 Proposed Method

37.2.1 System Structure

The proposed IoT system integrates IoT devices, depicted in Fig. 37.1, with strategically placed cameras across different areas of the field farm, employing a deep learning model.

Every thirty minutes, sensors on the camera nodes capture periodic snapshots of tomato leaves. Subsequently, each camera node wirelessly transmits the captured image data of tomato leaves to a central gateway station.

On a daily basis, the gateway utilizes Raspberry Pi Internet of Things devices to transmit the images to the processing center (server). Equipped with advanced computer vision technologies, the server works as the central hub for data computation and analysis. The core elements of the proposed IoT system include image processing and MLP-Mixer-based deep learning techniques. These components are employed to extract attributes from tomato leaves and promptly detect any signs of leaf diseases. Furthermore, the farmer receives timely notifications regarding the condition of the tomato crops.

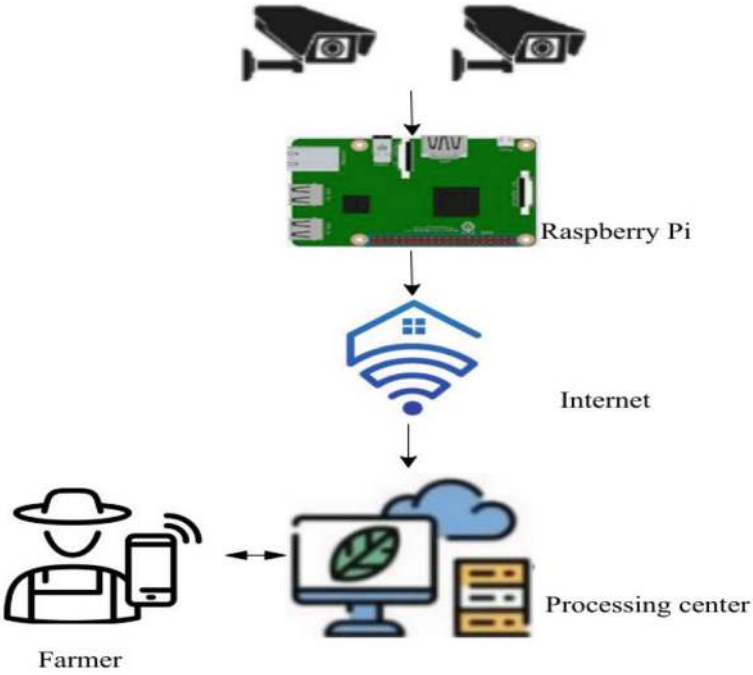


Fig. 37.1 IoT proposed system

37.2.2 MLP-Mixer Model

The MLP-Mixer comprises two types of layers: the channel-mixing MLP layer and the token-mixing MLP layer. In [12], the authors successfully replaced the MLP layer with the self-attention layer, yielding competitive results on image classification benchmarks.

Contrastingly, the Transformer encoder [13] alternates between multi-head self-attention and MLP blocks, with each block preceded by LayerNorm (LN).

The Multi-head Attention (MHA) is defined by the following formula:

$$U = X + \text{MHA}(\text{LayerNorm}(X), W_Q, W_K, W_V) \tag{37.1}$$

$$Y = (U^T + W_2\sigma(W_1\text{LayerNorm}(U)^T)^T \tag{37.2}$$

where X represents the path feature in the input, *LayerNorm* denotes the layer normalization, and W_V, W_Q, W_K, are the weights of learnable values V, queries Q, and keys K, respectively, with σ representing the activation function. The token-mixing block took the role of the self-attention in the mixer blocks, which are otherwise identical to the transformer block. The features X are projected along the channel dimension

by the channel-mixing block, displaying the following equation:

$$U = X + W_2\sigma(W_1\text{LayerNorm}(X)) \tag{37.3}$$

The output produced by the channel-mixing block U is fed into the token-mixing block:

$$Y = (U^T + W_4\sigma(W_3\text{LayerNorm}(U)^T)^T \tag{37.4}$$

Figure 37.2 illustrates the architecture of the proposed MLP-Mixer. The images with a given input dimension ($W \times H \times C$) are initially divided into S non-overlapping patches. The number of blocks is calculated as $H \times W/P \times P$, where (P, P, C) is the patch's dimension.

The feature maps of each patch from the original image are transformed into vectors, essentially converting the three-dimensional representation to two dimensions. The input to the stacked layers of the channel-mixing block and the token-mixing block, passed through the Gelu (Gaussian Error Linear Unit) activation function, results in the formation of the projected matrix. The channel-mixing MLP operates on the rows of this representation, while the token-mixing MLP operates on the columns. In other words, a mixing block facilitates communication across different patches on the same channel by blending tokens. Consequently, the number of channels remains constant, and the number of rows is expanded into the hidden dimension of the token layer.

The cross-entropy metric [14] measures the disparity between two probable distributions for a given random variable, serving as an evaluation and quantification tool for assessing the performance of the proposed model. The following equation describes the Binary Cross-Entropy:

$$L_{\text{BCE}} = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{37.5}$$

where $\hat{y}$ is the predicted value by the classifier. The last loss function is described in the following equation:

$$L_{\text{BCE}} = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

$$\text{Loss} = \frac{1}{N}\sum_{b=1}^{N} L_{\text{BCE}}(\text{x}^{\text{b}}) \tag{37.6}$$

where N is the number of patches.

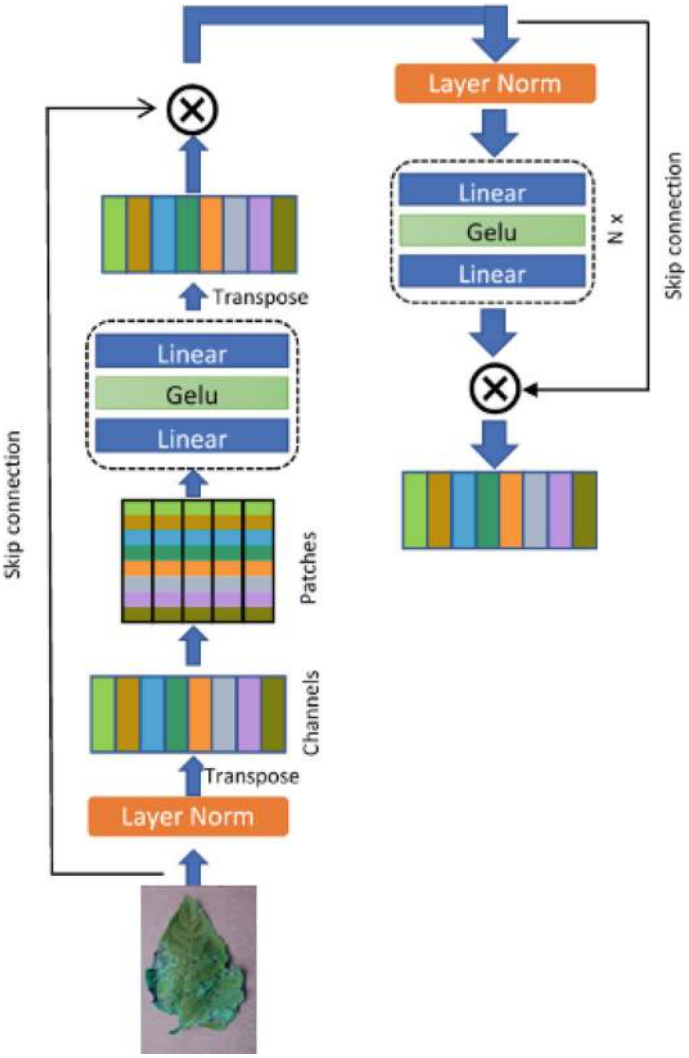


Fig. 37.2 MLP-mixer model architecture

37.3 Experimental Results

37.3.1 Dataset

In this experiment, we will utilize the PlantVillage Dataset benchmark, a substantial source of leaf image data for disease diagnosis. The dataset encompasses 54,309 recordings of both healthy and diseased leaves, spanning 14 distinct plant species.

Within this dataset, there are 18,160 images of tomato leaves categorized into nine classes for diseases and one for health. Figure 37.3 presents a sample from each class, showcasing leaves with varying degrees of disease involvement. The dataset provides comprehensive coverage of a wide spectrum of illnesses. Table 37.1 displays the number of samples for each class.

37.3.2 Implementation and Results

The computer system employed for this experiment featured an Intel (R) Core (TM) i7-7700 HQ 2.8 GHz processor and 16 GB RAM. The experiment was conducted using Python, utilizing TensorFlow packages, Scikit-learn, Keras, and OpenCV.

The evaluation of the proposed model for tomato disease diagnosis involved a diverse set of metrics, outlined as follows:

$$\text{Accuracy}(\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \tag{37.7}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{37.8}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{37.9}$$

$$\text{F1} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN}) \tag{37.10}$$

where the numbers for true positives, false negatives, false positives, and true negatives are indicated by the symbols TP, FN, FP, and TN, respectively.

The assessment of the proposed model utilized the PlantVillage Dataset, comprising 4585 images for testing and 18345 photos for training. The dataset encompasses labels categorized into nine classes: diseased tomato leaves and healthy plant leaves. For pre-processing, the dataset was initially resized to (64 × 64). Two separate experiments were conducted to compare the MLP-Mixer with transfer learning on one hand and other contemporary variations of deep learning algorithms based on MLP on the other.

The accuracy metric and parameter measures are employed. The models chosen for transfer learning include VGG16, AlexNet, VGG19 and FNet, a new perspective on MLP that Google researchers suggested [15].

Table 37.2 depicted the comparison results. Notably, while CNN and transformer models require quadratic costs, the MLP-Mixer exhibits linear computational complexity in terms of the number of input patches. The selection and optimization of parameters play a crucial role in determining the algorithm's effectiveness, particularly in adapting it to the specific dataset at hand.

As demonstrated in Table 37.2, the proposed model exhibits superior performance compared to both the transfer learning methods and MLP-based architecture of its

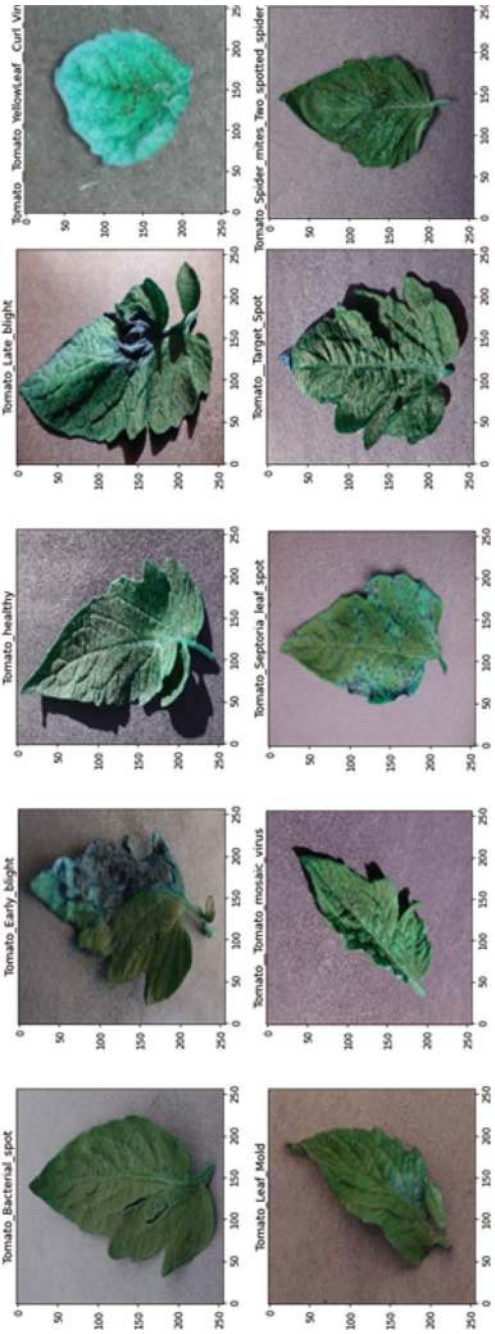


Fig. 37.3 Tomato leaves sample images from PlantVillage dataset

Table 37.1 PlantVillage tomato dataset

Class name	Samples	Augmentation	Trained samples	Tested samples
Tomato with bacterial spot	2127	2127	1702	425
Tomato with early blight	1000	2400	1920	480
Tomato with late blight	1909	2314	1851	463
Tomato with leaf mold	952	2352	1882	470
Tomato with Septoria leaf spot	1771	2181	1745	436
Tomato with two spotted spider mites	1676	2176	1741	435
Tomato with target spot	1404	2284	1827	457
Tomato with mosaic virus	373	2238	1790	448
Tomato with yellow leaf curl virus	5357	5357	1961	490
Healthy tomato	1591	2407	1926	481
Total number of images	16569	23429	18345	4585

Table 37.2 Comparison of the proposed model with the transfer learning techniques and Fnet

Techniques	Accuracy (%)	Total parameters	Trainable parameters
VGG16	90.47	14,965,578	250,890
AlexNet	95.43	58,322,314	58,322,314
VGG19	85,45	20,106,314	81,930
Fnet	97.80	715,025	715,018
Proposed MLP-mixer	98.34	261,585	261,578

counterparts. In this experiment, the MLP-based architecture achieved a validation accuracy of 97.80% (Fnet) and 98.34% (MLP-Mixer), surpassing the accuracy of the evaluated transfer learning models 90.47% (VGG16), 95.43% (AlexNet), and 85.45% (VGG19).

Furthermore, it can be argued that MLP-based architectures excel in both parameter efficiency and performance accuracy when compared to transfer learning models. The MLP-Mixer, with the fewest parameters, outperforms the transfer learning models, especially considering AlexNet with the highest number of trainable parameters. Consequently, in terms of efficiency and execution time, the MLP-Mixer can be considered the superior choice. Table 37.3 presents a comparison of existing models from the literature. The Precision, Recall and F1-score metric of MLP-mixer for the ten classes were depicted on Table 37.4.

Table 37.3 Comparative analysis of existing models

References	Techniques	Classification accuracy (%)
[7]	PCA–whale optimization-based deep learning model	94
[16]	Random forests classifier	94
[17]	Neural network classifier	97.30
[18]	Optimized MLP	91.45
[19]	MobileNet V2	95.62

Table 37.4 Performance metrics for each class

Class name	Precision	Recall	F1-score	Support
Tomato with bacterial spot	0.98	0.98	0.98	424
Tomato with early blight	0.99	0.97	0.98	490
Tomato with late blight	0.96	1.00	0.98	444
Tomato with leaf mold	1.00	0.99	0.99	473
Tomato with Septoria leaf spot	0.99	0.99	0.99	437
Tomato with two spotted spider mites	0.98	0.98	0.98	436
Tomato with target spot	0.95	0.98	0.96	442
Tomato with mosaic virus	1.00	0.98	0.99	454
Tomato with yellow leaf curl virus	0.99	0.98	0.99	496
Healthy tomato	1.00	0.98	0.99	489

37.4 Conclusion

This study aimed to develop and evaluate an Internet of Things (IoT) system for the accurate identification of nine tomato leaf diseases and distinguish them from healthy leaves. The evaluation of the proposed MLP-Mixer model was conducted in comparison with several well-established transfer learning methods, including VGG16, AlexNet, VGG19, and FNet. The results highlight the effectiveness and competitiveness of the MLP-Mixer model.

The context of rising global population, predicted to reach a 29.33% increase by 2050, has necessitated a 60% increase in food production, as highlighted in [20]. Traditional agricultural methods are deemed insufficient to meet this growing demand, prompting a search for innovative solutions by agro-businesses and farmers. Deep learning models leveraging advanced technologies emerge as promising solutions to address the challenges in agriculture, offering the potential to enhance productivity and reduce waste. In our future work, we will extend the presented IoT system to monitor the environmental factors such as the temperature, the humidity and the soil pH to evaluate the health of the plants and to get the maximum yield.

Acknowledgements The authors acknowledge the support of King Saud University, Riyadh, Saudi Arabia.

References

1. FAO.: New standards to curb the global spread of plant pests and diseases. https://www.fao.org/news/story/en/item/1187738/icode/. Accessed 17 Mar. 2022
2. Zhao, S., Peng, Y., Liu, J., Wu, S.: Tomato leaf disease diagnosis based on improved convolution neural network by attention module. Agriculture **11**(7), 651 (2021)
3. Rabhi, B., Elbaati, A., Boubaker, H., Hamdi, Y., Hussain, A., Alimi, A.M.: Multi-lingual character handwriting framework based on an integrated deep learning based sequence-to-sequence attention model. Memetic Comput. **13**, 459–475 (2021). https://doi.org/10.1007/s12293-021-00345-6
4. Rabhi, B., Elbaati, A., Boubaker, H., Pal, U., Alimi, A.M.: Multi-lingual handwriting recovery framework based on convolutional denoising autoencoder with attention model. Multimedia Tools Appl. **1–32**,(2023). https://doi.org/10.1007/s11042-023-16499-z
5. Atila, Ü., Uçar, M., Akyol, K., Uçar, E.: Plant leaf disease classification using EfficientNet deep learning model. Eco. Inform. **61**, 101182 (2021)
6. Ahmed, S., Hasan, M.B., Ahmed, T., Sony, M.R.K., Kabir, M.H.: Less is more: lighter and faster deep neural architecture for tomato leaf disease classification. IEEE Access **10**, 68868–68884 (2022)
7. Gadekallu, T.R., Rajput, D.S., Reddy, M.P.K., Lakshmanna, K., Bhattacharya, S., Singh, S., Alazab, M.: A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. J. Real-Time Image Proc. **18**, 1383–1396 (2021)
8. Dhahri, H., Rabhi, B., Chelbi, S., Almutiry, O., Mahmood, A., Alimi, A.M.: Automatic detection of COVID-19 using a stacked denoising convolutional autoencoder. Comput. Mater. Contin. **69**(3) (2021)
9. Rabhi, B., Elbaati, A., Hamdani, T.M., Alimi, A.M.: ASAR 2021 competition on online signal restoration using arabic handwriting Dhad dataset. In: Document Analysis and Recognition-ICDAR 2021 Workshops: Lausanne, Switzerland, Proceedings, Part I 16, pp. 366–378. Springer International Publishing. (2021)
10. Bouzidi, S., Imen, J., Alimi, M.A.: A vision transformer approach with L2 regularization for sustainable fashion classification. SSRN 4686032
11. Hamdi, Y., Boubaker, H., Rabhi, B., Qahtani, A.M., Alharithi, F.S., Almutiry, O., Alimi, A.M.: Deep learned BLSTM for online handwriting modeling simulating the Beta-Elliptic approach. Eng. Sci. Technol. Int. J. **35**, 101215 (2022)
12. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Dosovitskiy, A.: MLP-mixer: an all-MLP architecture for vision (2021). arXiv:2105.01601.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
14. Yi-de, M., Qing, L., Zhi-Bai, Q.: Automated image segmentation using improved PCNN model based on cross-entropy. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., pp. 743–746. IEEE (2004)
15. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms (2021). arXiv:2105.03824
16. Basavaiah, J., Arlene Anthony, A.: Tomato leaf disease classification using multiple feature extraction techniques. Wirel. Pers. Commun. **115**(1), 633–651 (2020)
17. Sannakki, S.S., Rajpurohit, V.S.: Classification of pomegranate diseases based on back propagation neural network. Int. Res. J. Eng. Technol. (IRJET) **2**(2) (2015)
18. Vamsidhar, E., Rani, P.J., Babu, K.R.: Plant disease identification and classification using image processing. Int. J. Eng. Adv. Technol **8**(3), 442–446 (2019)
19. Zaki, S.Z.M., Zulkifley, M.A., Stofa, M.M., Kamari, N.A.M., Mohamed, N.A.: Classification of tomato leaf diseases using MobileNet v2. IAES Int. J. Artif. Intell. **9**(2), 290 (2020)
20. World Population Prospects 2019: highlights. https://www.un.org/en/sections/issues-depth/population/. Accessed 2 Dec. 2019

Chapter 38
Experiments on Semantic Segmentation of Medical Images with Multilabels

Ana-Maria Bumbu and Anca Ignat

Abstract The current work represents a study of different methods used in developing convolutional neural network models for semantic segmentation or pixel classification tasks when dealing with datasets that have noisy labels. We perform some comparisons between different methods that treat multilabel pixel annotation of medical images. We describe the preprocessing step of the dataset and how to obtain agreement masks for the training of deep neural networks. We test these methods on Gleason2019, a dataset of medical images of Prostate Cancer tissue microarray (TMA). We study the task of pixel-level classification of Gleason grade. The scope is to see how our classification models behave when trained on different variations of the computed masks. We also analyze which is the best method to handle the noise found in the annotations of these images.

38.1 Introduction

Deep learning tasks demand as the first step of development a large dataset to experiment with. Either we talk about image segmentation and classification or detection as supervised machine learning tasks, all the data need to contain labels. Depending on the task performed there are different kinds of labels. For image classification, this consists of a tag associated with every image with the number of unique tags equal with the classes that the model can classify. In segmentation tasks the labels can be seen as an image with the same size as the target and containing information of the class to which each pixel in the image belongs.

Data labelling is an indispensable step in data preprocessing and there are different ways to be performed, depending on complexity and the data size and also, the costs, time and resources allocated for this process [1, 2]. The best way of doing this step is by humans, preferably experts in domain, which can be expensive. Another way to label data is with automated systems developed to get rid of manual labelling,

A.-M. Bumbu · A. Ignat (✉)
Faculty of Computer Science, University "Alexandru Ioan Cuza" of Iaşi, Berthelot str. 16, Iaşi 700483, Romania
e-mail: ignata@uaic.ro

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_38

however, this type of approach can have a negative impact on the quality of the results. However, all these methods can introduce different levels of noise in the labels. When we talk about label noise we refer to different deviations, corruptions or imperfections from the true dataset. The noise can affect the training, behavior and performance of deep learning models in a negative way depending on his type. In [3] label noise is categorized in three types: "uniform noise", "class-dependent noise" and "feature dependent noise". This classification is made accordingly with the probability of a label to be misclassified. Therefore, as its name says, for uniform noise this probability is distributed equal with the number of classes, for "class-dependent" the probability hangs on the class similarities and for "feature dependent noise" the probability is modeled based on the similarities of the features found in samples.

There are a lot of studies that are trying to show and combat this negative impact of label noise during the training of deep neural networks [4]. In [5] it is mentioned that training the convolutional neural networks in the presence of a high rate of noise leads to the phenomenon of over-fitting and this makes the model generalize poorly and reduce its performances. Also, it was shown in [6] that between the prediction errors and level of noise is a "positive correlation". This fact makes experts say that neural networks can learn suitable information from the datasets which contain noise. The robustness of these neural network models trained in the presence of label noise is dependent on the noise distribution, thus, these are more sensitive to locally concentrated noise than in case the data contains noise uniformly dispersed. In [7] the authors study the learning dynamics and observe when the clean pixels are fitted.

There are many approaches about training networks on noisy datasets. The simplest one consists of a preprocessing stage added in the flow of the model. This means identifying incorrect labels and fixing or removing them from the dataset before training step or in parallel with it. Nevertheless, such an approach can be impractical for bigger datasets and can produce errors like removing important samples. In [8] the authors use a network, CleanNet, for cleaning the datasets like. CleanNet compares the "query embedding" of a sample with its "class embedding" in order to predict if that sample has the wrong annotation. In [9] the authors propose an iterative self-learning framework that improves CleanNet by using multi-embedding for class representations which produced a higher classification accuracy. Also, they removed the clean dataset needed by CleanNet using an iterative framework for estimating the correct labels. In [10] a "data-driven teacher-student architecture" is introduced in order to produce the so-called "easy-to-hard curriculum weights". A method involving two CNN models that train each other is called co-teaching and is a good approach to deep learning with noisy labels problems [11]. At "co-teaching", firstly, in every mini-batch, the models filter noisy annotations regarding their "memorization effects". After that, the network teaches the remaining images with a small loss on how to update the parameters. Finally, there will be two networks with different abilities to filter the errors that occur during training. In [2] the authors use two coupled CNN that learn together, using only data with noisy labels, the trustworthiness of annotators and the agreement of experts.

Some approaches change the focus from the noisy datasets to the networks, trying to improve the training procedures, and parameters or develop new architectures that can handle labels. In this regard, some studies designed a new layer which can match the network outputs to the distribution of noise found in labels. The "noise layer" should be added at the last part of the neural network [12]. Another subject that was analyzed in the context of noisy labels is the loss functions robustness [13, 14].

Another way of dealing with noisy labels is by reweighting algorithms which intend to give lower weights to the images with annotation that contain noise from the dataset. The authors of [15] used a meta-learning technique to weight the training data. The method proposed in [16] requires a clean subset with clean labels beside the main noisy dataset, and exploits correlations found between the noisy samples and correctly labelled ones from the "auxiliary" dataset. Another mode of using reweighting was proposed by [17] where a re-weighting scheme should be learned from the data samples. The authors employed a multilayer perceptron that contains in its architecture one hidden layer to learn how to properly weigh the noisy samples from the dataset. The training of this architecture also requires a clean, small subset of images. In [18] the correction of the noisy labels is performed by using image features based on a region-scalable fitting.

In our present work, we compute agreement masks based on three methods (mean SIMPLE and STAPLE) and test their efficacy using two adapted U-Net architectures.

Our work is structured in five sections. The second section presents the dataset and the preprocessing operations. How the agreement masks are computed is described in the third section. The fourth section is dedicated to experiments and results. The final section is dedicated to conclusions.

38.2 Dataset and Preprocessing Operations

In the following, we'll describe the collection of images used in the experiments and the preprocessing procedures applied.

38.2.1 Dataset

Prostate cancer (PCa) is considered the second most common cancer and one of the deadliest cancers among men around the world. To get a diagnosis, the pathologist has to examine the cellular and morphological patterns and based on them to rank the level of aggressiveness of PCa according to the Gleason scoring system. This Gleason system is based exclusively on the architectural pattern of the tumor and consists of grade values from 1 to 6. The grades 1, 2, and 6 are for benign tumors and grades 3, 4 and 5 signal out different cancerous levels of tissue regions. The sum of the first and second dominant grades is the final Gleason score. This assessment of Gleason

score is time-consuming for pathologists and suffers from very high inter-observer variability because it relies on the experience of the pathologist.

In this paper, we used the Gleason 2019 dataset [19], part of the "Grand Challenge for Pathology at MCCAI 2019", for our task of semantic segmentation (pixel classification) from images with noisy labels. This work focuses on the prediction of pixel-level Gleason grade testing different methods to improve the score, using the multi-label information from the dataset.

The dataset contains images of tissue microarray (TMA) about prostate cancer ranked by six pathologists with experience between 1 and 27 years according to the Gleason scoring system. Each pixel from the TMA image is annotated with one of the numbers 0, 1, 3, 4, 5, and 6 in detail by the expert. The dataset contains a total number of 244 images for training, these images having a huge size of 5120×5120. We only use the training set because the labels for the test set haven't been released yet. In many cases, the annotations for an image are quite different depending on the pathologist. On the same image, one pathologist can indicate that all tissue is benign and other pathologists identify cells with Gleason grade 3 or 4. This phenomenon is presented in Fig. 38.1.

Besides the noise found in labels, another issue with the data provided is the existing class unbalance. We observed that in this dataset not every image has annotations from all the six pathologists and this is also an issue for the computation of an accurate annotation. As can be seen in Table 38.1, only pathologist 5 labelled all the existing images, pathologists 1, 3 and 4 did not annotate just a few images, pathologist 2 annotated a little over half of the images and the last pathologist annotated the least number of images. Table 38.1 also shows the average percentage of

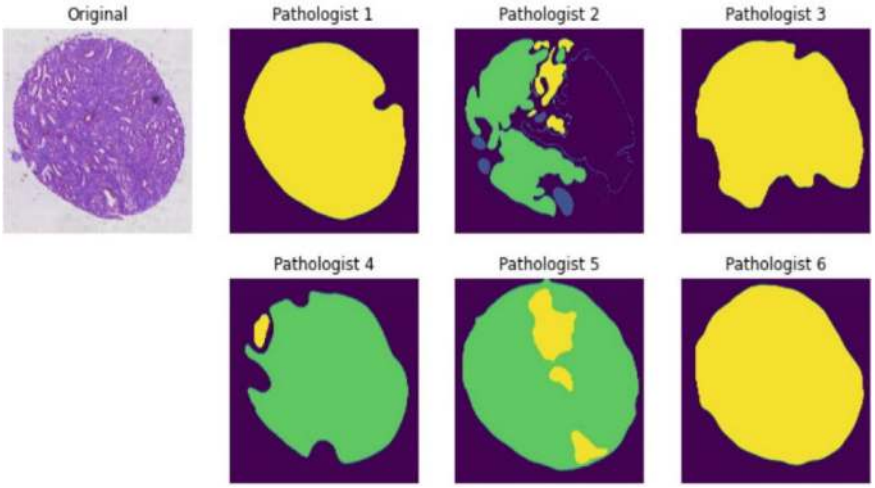


Fig. 38.1 Pathologists' labels for the same image: label 0—background/indigo pixels, label 1—blue pixels, label 3—green pixels, label 4—yellow pixels

Table 38.1 Number of annotated images per pathologist and average percent of annotated labels

Pathologist	1	2	3	4	5	6
No. of annotated images	242	139	240	241	244	65
Average background pixels (%)	52.52	76.90	57.78	50.60	44.96	42.58
Average label 1 (%)	1.55	5.71	5.56	10.32	7.19	11.30
Average label 3 (%)	10.32	5.50	14.33	14.91	21.61	17.83
Average label 4 (%)	28.37	7.01	22.25	23.83	25.51	23.52
Average label 5 (%)	2.62	4.00	0.08	0.35	0.73	4.50
Average label 6 (%)	4.62	0.89	0	0	0	0

pixels annotated with a certain label. There are significant differences in the labelling process among the six pathologists.

The class distribution of each label is also unbalanced, the images with the class of Gleason 5 are the fewest 4,1%, benign class are the most images 53,7%, and Gleason 3 are 25,4% while Gleason 4 are just 16,8%. These percentages are calculated as the mean of all existing annotations for each image of the dataset.

38.2.2 Data Preprocessing

The high size of images is the biggest impediment in working with neural networks because it requires large computational resources and is time-consuming.

Therefore, in the first attempt of creating a training dataset all the images and masks are resized from 5120 × 5120 to size 640 × 640 using Nearest Neighbor interpolation because it is the quickest interpolation procedure and it preserves the original content of the images. To increase the number of images an augmentation step was applied starting with the rotation of all images with an angle chosen at random from 90, 180 or 270°. Next, a vertical flip was applied to the initial images, after that a horizontal flip and a vertical one were applied to the images rotated before. The resulting dataset contains a total of 976 images with the dimension of 640 × 640 × 3 and also the corresponding 976 annotations of size 640 × 640 × 1. Regarding the existence of multiple annotations per image, the first step was to compute a single agreement mask that is used as ground truth information. When computing the mean of all the existing annotations the final results were rounded to the nearest integer value, thus we keep the original class values (0, 1, 3, 4, 5 and 6).

38.3 Computing Agreement Masks

Computing the mean of all annotations to create an agreement annotation for each image can be seen as a naive method because some information can lead to a wrong diagnosis. Hence, an alternative to construct these agreement labels was considered. We used two methods for computing the agreement mask: one was the SIMPLE (Selective and Iterative Method for Performance Level Estimation) and STAPLE (Simultaneous Truth and Performance Level Estimation) to fuse the multi-label information [20–23]. STAPLE is a weighted voting algorithm composed of an iterative process where in the first step the filter performs a pixel-wise combination of the images into an input segmentation and the second step consists of the estimation of the accuracy of each pathologist compared to this initial segmentation. This cycle of estimating the pathologist's accuracy and recombining the input segmentation by weighting the votes of each pathologist according to their accuracy will repeat until the test segmentation converges. This algorithm was applied to each grouping of annotations for the original 244 images. After that, the augmentation step was applied and resulted in the second dataset with a total of 976 images.

In the same manner, another dataset was created, this time using the SIMPLE algorithm to create the agreement annotations. This method is related to STAPLE meaning that it also consists of an iterative process of estimating the accuracy of the input segmentations and the ground truth segmentation. Thus, a major difference between these two algorithms exists, because at every step the input segmentations that have the worst accuracy are discarded by the SIMPLE algorithm. Moreover, the SIMPLE algorithm, as the name suggests, is simpler and also faster because it is not based on the expectation-maximization algorithm in its structure. Thus, this new dataset created using the SIMPLE method contains also 976 images obtained after the augmentation step. The images below show the difference between the first three methods used to compute the agreement labels from the annotations provided by the six pathologists (Fig. 38.2).

All these processed variants of the initial Gleason2019 dataset will be exploited in the experiments described in the next section.

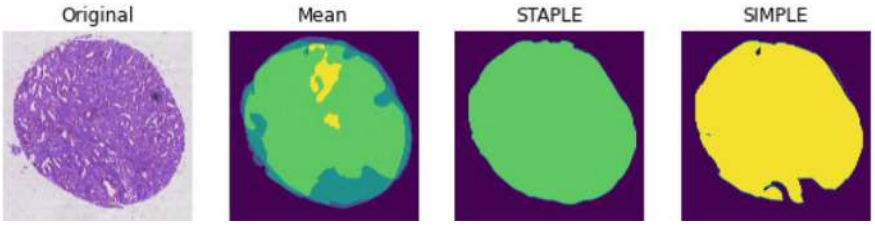


Fig. 38.2 Resulted agreement labels

38.4 Experiments and Results

The architectures used to compare the datasets presented in the chapter above are variations of the U-Net model [24]. The dataset used in the first experiment is the augmented dataset that contains 976 images with dimensions 640 × 640 × 3.

38.4.1 Experiment 1

In this experiment, the architecture of U-Net is inspired by the Xception network [25] because it uses depth-wise separable convolutional layers. The network is made from two parts: the down-sampling part and the up-sampling one. The down-sampling part contains four convolutional blocks, the first block is the smallest with one convolutional layer containing a kernel with size 3 × 3 and 32 filters, followed by a layer for batch normalization and one *ReLU* activation layer. In the next three blocks, filters of dimensions 64, 128 and 256 were used. Each of these contains two separable convolutional layers also followed by batch normalization and *ReLU* activation layers. After that, max-pooling is applied to increase the size of the receptive field and decrease the size of the feature map. At the end, a residual convolutional layer was applied as the first layer of the block and then the last layer was added.

In the up-sampling part, four convolutional blocks with 256, 128, 64 and 32 filters were used to up-sample the feature map back into the initial size of inputs. Each block contains two transpose convolutional layers followed by a batch normalization layer together with a *ReLU* activation layer, an up-sampling layer and a residual part. This residual part is composed of another up-sampling layer and a simple convolutional layer which are concatenated with the previous block through the first layer. At the end of the network a convolutional layer which contains six filters, a kernel size of 1 × 1 and the softmax function as activation was applied to get the probability map of each pixel. As hyper-parameters, Adam was used for the optimization of the network, the sparse categorical cross-entropy as loss function, the training lasted 30 epochs, and the size of the batch was 8.

38.4.2 Experiment 2

The U-net architecture used in this experiment has as the encoder part a pre-trained MobileNetV2 network because it has good results in learning robust features and also decreases the size of the network trainable parameters. The encoding part is made from specific outputs of the intermediate layers that exist in the model architecture. The MobileNetV2 [26], proposed by Google, is an improved version of MobileNetV1 a network with a reduced complexity, cost and model size, suitable for mobile devices and many other machines with low computational power. In this second version an

improved module with "inverted residual structure" is introduced and one more improvement is that the non-linearities existing in the narrow layers are removed. The classical residual structures have in the input some expanded representations, but the MobileNetV2 model contains lightweight depth-wise convolutions to filter the features found in the intermediate expansion layer. Therefore, the encoder will be composed of the output of the pre-trained layers of the MobileNetV2 model.

The decoder part of this architecture consists of four up-sampling blocks with the size of filters 32, 64, 128 respectively 256, which were initialized as a sequential model having the first layer a transpose convolutional one, the next one is a batch normalization layer followed by a *ReLU* activation layer. To produce the U shape of the network, the layers between the pre-trained encoder and decoder are concatenated together. At the end of this architecture, there are two layers used for the classification of pixels: one transpose convolution and one normal convolution with six filters corresponding to each class.

38.4.3 Results

The results of these experiments are evaluated using two common metrics employed in semantic segmentation tasks: Intersection over Union and Dice coefficient. Consider the ground-truth information for a certain label/class. Using the models calculated in each experiment, one computes for each pixel if it belongs to that class or not. With this label assignment, one can compute the TP, FP, FN, TN—true positive, false positive, false negative, respectively true negative values. To evaluate the quality of the segmentation task the following formulae for Intersection over Union and Dice coefficient evaluators are employed:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \ \text{Dice} = \frac{2\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{38.1}$$

We use in Table 38.2 the mean value for the evaluators in (38.1) over all the five possible labels.

In the first experiment the best results were provided by the STAPLE algorithm for computing the agreement mask and for the second experiment best results were computed for the SIMPLE algorithm although the STAPLE results were close.

Table 38.2 Mean IoU and Dice coefficient results for the two experiments

	Experiment 1		Experiment 2	
Dataset	mIoU	Dice	mIoU	Dice
Mean (%)	61.5	55.0	58.2	49.3
Staple (%)	78.2	64.1	68.7	61.6
Simple (%)	72.2	63.8	69.3	59.1

Figure 38.3 shows some test samples of the predicted labels for the two experiments.

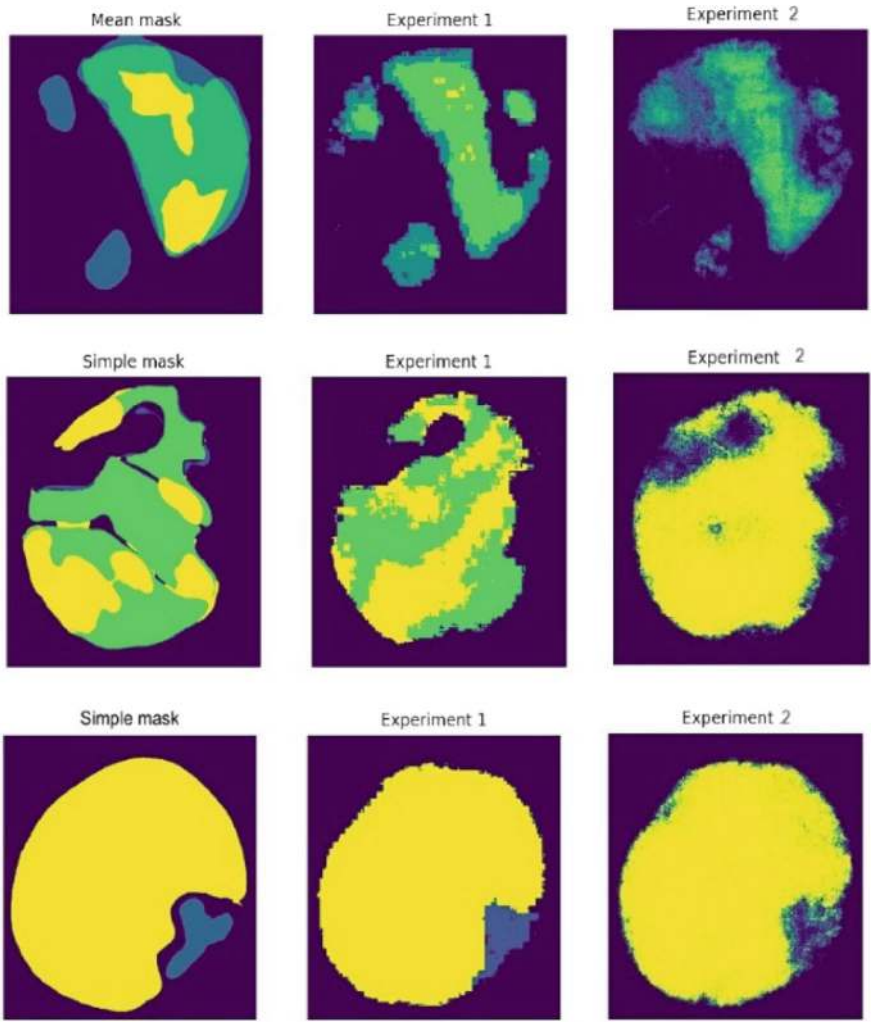


Fig. 38.3 Prediction of experiments

38.5 Conclusions

The current paper addresses the problem of semantic segmentation of medical domain images that contain multiple labels in their annotations. In this work we present and compare different methods of noisy label processing aiming to improve the learning of convolutional neural network models. For the practical part, a representative set of images was chosen for the presented theme, the Gleason 2019 dataset. We used mean, SIMPLE and STAPLE multi-label fusion methods to model the annotations of this dataset. The dataset was augmented with rotations and flip operations. Two experiments were performed using variations of the well-known U-net model. The results show that STAPLE and SIMPLE methods provide better results than the mean algorithm.

One of the future directions of research is to apply the re-weighting technique on the noisy samples during training together with transfer learning via fine-tuning to improve the training of the neural networks in the presence of noise.

References

1. Gurari, D., Theriault, D., Sameki, M., Isenberg, B., Pham, T.A., Purwada, A., Solski, P., Walker, M., Zhang, C., Wong, J.Y., et al.: How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1169–1176 (2015)
2. Zhang, L., Tanno, R., Xu, M., Huang, Y., Bronik, K., Jin, C., Alexander, D.C.: Learning from multiple annotators for medical image segmentation. Pattern Recognit. **138** (2023)
3. Görkem, A., Ulusoy, İ.: Label noise types and their effects on deep learning (2020). arXiv: 2003.10471
4. Li, D., Wang, C.: In-depth research and analysis of multilabel learning algorithm. J. Sens. (2022)
5. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning, pp. 233–242, PMLR (2017)
6. Drory, A., Avidan, S., Giryes, R.: On the resistance of neural nets to label noise (2018). arXiv: 1803.11410
7. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2606–2616 (2022)
8. Lee, K.-H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5447–5456 (2018)
9. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels (2019). arXiv:1908.02160
10. Jiang, L., Zhou, Z., Leung, T., Li, L.-J., Fei-Fei, L.: Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels (2018). arXiv:1712.05055v2
11. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: NeurIPS (2018)
12. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: Training convolutional networks with noisy labels (2014). arXiv:1406.2080

13. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) (2017)
14. Wang, X., Kodirov, E., Hua, Y., Robertson, N.M.: Improving MAE against CCE under label noise (2019). arXiv:1903.12141
15. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning (2018). arXiv:1803.09050
16. Azadi, S., Feng, J., Jegelka, S., Darrell, T.: Auxiliary image regularization for deep cnns with noisy labels (2015). arXiv:1511.07069
17. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Metaweight-net: learning an explicit mapping for sample weighting (2019). arXiv:1902.07379
18. Liu, S., Li, Y., Chai, Q.W., Zheng, W.: Region-scalable fitting-assisted medical image segmentation with noisy labels. Expert Syst. Appl. **238** (2024)
19. https://gleason2019.grand-challenge.org/. Accessed 4 Feb. 2024
20. Warfield, S., Zou, K., Wells, W.: Validation of image segmentation and expert quality with an expectation-maximization algorithm. In: MICCAI 2002, pp. 298–306 (2002)
21. Warfield, S., Zou, K., Wells W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004)
22. Rohlfing, T., Russakoff, D.B., Maurer, C.R., Jr.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging **23**, 983–994 (2004)
23. Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging **29**(12), 2000–2008 (2010)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 9351, pp. 234–241 (2015)
25. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: inverted residuals and linear bottlenecks. In The IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

Chapter 39
Named Entity Recognition for Algerian Arabic Dialect Using Multi-dialect-Arabic-BERT Based Architectures

Manel Affi and Chiraz Latiri

Abstract Named Entity Recognition represents a crucial component in natural language processing (NLP), involving the identification and categorization of named entities, such as persons, organizations, and locations, within textual data. Despite significant advancements in NER for prominent languages like English and Arabic, recognizing named entities within specific dialects presents distinctive challenges. Algerian Arabic, a widely spoken variant with unique linguistic characteristics, has garnered limited attention in NER research compared to standardized Arabic varieties. This study introduces a novel architecture that addresses the NER challenges specific to Algerian Arabic. Our proposed framework integrates the Multi-dialect BERT model, Bidirectional Gated Recurrent Units (Bi-GRU), and Conditional Random Field (CRF) layer. Our experimental findings showcase promising results across all architecture variants. Particularly, the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model emerges as the most effective, achieving a remarkable 93.04% F1-score. This success underscores its proficiency in capturing sequential dependencies and decoding complex linguistic structures inherent in Algerian Arabic text.

39.1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and classifying named entities within text into predefined categories such as persons, organizations, locations, dates, and numerical expressions. The accurate identification of named entities plays a crucial role in various NLP applications, including information extraction, question answering, and sentiment analysis.

M. Affi (✉)
RIADI, ENSI, University of Manouba, Tunis, Tunisia
e-mail: manel.affi@ensi-uma.tn

C. Latiri
LIPAH, FST, University of Tunis El Manar, Tunis, Tunisia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7_39

While NER has been extensively studied for major languages such as English and Arabic, with significant progress achieved through various approaches and models, the recognition of named entities in specific dialects presents unique challenges. Arabic, as a language with diverse dialects across different regions, poses particular difficulties due to variations in syntax, vocabulary, and grammatical structures.

Among the Arabic dialects, Algerian Arabic stands out as one of the most widely spoken and culturally significant variants. Despite its importance, Algerian Arabic has received relatively limited attention in NER research compared to standardized Arabic varieties. The Algerian dialect poses unique linguistic characteristics and data scarcity, making NER tasks a field that remains relatively underexplored. According to our investigation, the challenge in Algerian dialect NER research lies in the scarcity of resources for developing machine learning or deep learning tools and for conducting evaluations [1]. Consequently, current efforts in Algerian NER primarily revolve around corpus construction. NERDz, an Algerian NER dataset, was introduced by [2] as an Algerian NER dataset, as an extension of the NArabizi treebank. NERDz covers eight categories of entities, including PER, GPE, ORG, NORP, EVT, LOC, PROD, and MISC. On a similar note, paper [3] presented DzNER, an Algerian NER dataset consisting of over 21,000 sentences sourced from Algerian social media platforms and annotated manually for PER, ORG, and LOC entities. Additionally, paper [4] investigated the impact of segmentation and the use of Latin characters in the Algerian dialect NER, employing pre-trained models such as AraBERT, MARBERT, ARBERT, DziriBERT, and mBERT on a newly annotated dataset. Their findings revealed that ARBERT achieved the highest performance in Arabic characters, with an F1-score of 81.9% on segmented data and 84.4% on unsegmented data, while mBERT excelled in Latin characters with an F1-score of 67.6%.

In this paper, we address the challenge of NER in Algerian Arabic by proposing a novel architecture based on the Multi-dialect BERT model, Bidirectional Gated Recurrent Units (Bi-GRU), and Conditional Random Field (CRF) layer. Our approach leverages the rich contextual embeddings provided by Multi-dialect BERT to capture the nuances of Algerian Arabic text. The Bi-GRU layer enhances the model's ability to learn long-range dependencies and contextual information, while the CRF layer enables joint decoding to improve the coherence of named entity sequences.

The primary objective of this paper can be outlined as follows:

- To investigate the effectiveness of proposed architectures in accurately identifying named entities in the Algerian dialect.
- Evaluate the performance of different architectures, including:
  - Fine-tuning the Multi-dialect-Arabic-BERT model.
  - Concatenating Multi-dialect-Arabic-BERT with Bi-GRU.
  - Incorporating Multi-dialect-Arabic-BERT with CRF.
  - Integrating Multi-dialect-Arabic-BERT, Bi-GRU, and CRF layers.

- Determine the best architecture through comparative analysis.
- Perform error analysis on the best models to identify areas for improvement and enhance model performance.

39.2 Proposed Model

In this section, we outline the design of our neural network model leveraging multi-dialect BERT word embeddings. The proposed system consists of three core stages: the word embedding layer, the context layer, and the tag decoder layer.

As depicted in Fig. 39.1, our proposed model comprises three main blocks, elucidated below.

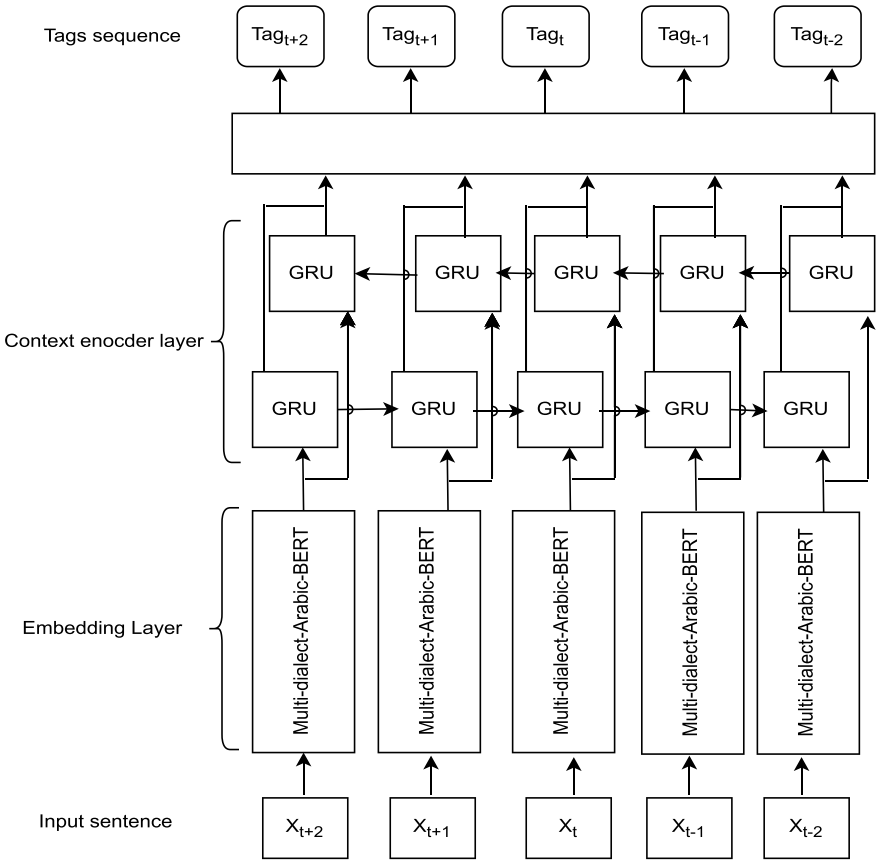


Fig. 39.1 Principal model architecture

39.2.1 Embedding Layer

The feature extraction layer is crafted to furnish the primary attributes essential for our NER system. The efficacy of these features significantly influences the model's performance. Traditionally, feature engineering involves manual crafting based on specific rules that may not always be pertinent across different domains. Consequently, modern NLP approaches often employ diverse architectures of deep neural networks to produce distributed word representations, enabling them to capture both syntactic and semantic patterns in words. With distributed embeddings, the model becomes more adaptable, as each word corresponds to dense, low-dimensional real-valued vectors in space. This means that words with similar semantic and syntactic properties will have comparable vector representations. However, generating high-quality word vectors can pose challenges. Ideally, these vectors should accurately capture the complex characteristics of word usage and how such usage varies based on linguistic context.

In recent years, various tools like word2vec [5] and GloVe [6] have become prevalent in the NLP domain. To attain superior word embeddings, researchers have introduced novel techniques to generate diverse embeddings for the same word based on its context [7, 8]. In the realm of distributed embeddings, words are mapped to vectors within a continuous space, enabling the capture of both syntactic and semantic relationships among them. In our study, we leverage the capabilities of BERT for Arabic language processing, we utilize publicly accessible pre-trained resources tailored for research applications, namely, the Multi-dialect-Arabic-BERT model [9] to serve as our foundational framework for distributed word embeddings. Initially, the weights for Multi-dialect-Arabic-BERT were initialized from Arabic-BERT. Following this, the model underwent further training using a dataset containing 10 million Arabic tweets sourced from the unlabeled data provided by The Nuanced Arabic Dialect Identification (NADI) shared task.[1]

39.2.2 Context Encoder Layer: Bi-GRU

The context layer is designed to capture local dependencies by considering neighboring words for each word in the input sequence. Local context plays a critical role in accurately predicting labels, as neighboring tokens within a sentence often exhibit strong relationships. Thus, modeling the local context information for each word is essential in the NER task.

In NER, recurrent neural networks (RNNs) are commonly used as context encoder models. RNNs process input sequences sequentially, and variations such as GRU have been introduced to mitigate the vanishing gradient problem [10]. The vanishing gradient problem arises when RNNs struggle to remember information from distant time steps, particularly in longer input sequences. To tackle this challenge and

[1] https://sites.google.com/view/nadi-shared-task.

effectively utilize both past and future information for prediction in the NER task, as proposed by [11], we utilize Bidirectional GRU (Bi-GRU) networks as the context encoder layer.

In our approach, we employ a Bi-GRU to process the input sentence both forward (Eq. 39.1) and backward (Eq. 39.2). The hidden states from both directions are then concatenated to form the final hidden state (Eq. 39.3).

$$\overrightarrow{h_t} = g(\overrightarrow{h_{t-1}}, [X_t^{emb}]) \tag{39.1}$$

$$\overleftarrow{h_t} = g(\overleftarrow{h_{t+1}}, [X_t^{emb}]) \tag{39.2}$$

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{39.3}$$

Here, $g(.)$ represents the unidirectional GRU unit, $\overleftarrow{h_t}$ and $\overrightarrow{h_t}$ denote the hidden states of the backward and forward GRU units, respectively, and X_t^{emb} signifies the embeddings acquired from the Multi-dialect-Arabic-BERT model at the time instance t.

39.2.3 Tag Encoder Layer: CRF Layer

For sequence labeling tasks, it is advantageous to consider the interdependence among tags in local neighborhoods. When analyzing an input sentence, jointly decoding the label chain that produces the optimal label sequence is crucial. In NER tasks using the IOB annotation scheme, there exists a significant correlation between labels; for instance, I-PER cannot follow I-ORG. Thus, modeling label correlations is vital for the NER task [12, 13]. Following the approaches of [12, 14], we integrate a CRF [15] layer on top of the context layer to collectively capture label dependencies.

39.3 Training Mechanism

39.3.1 Datasets

To evaluate the effectiveness of our proposed methodology, we conducted comprehensive experiments using a widely recognized benchmark dataset for Algerian dialect NER, namely, the DzNER-Corpus [3]. The dzNER dataset consists of more than 21,000 sentences, equivalent to over 220,000 tokens collected from Algerian Facebook pages and YouTube channels. The annotation process involved manual annotation by two professional annotators, focusing on three entity types: PER (person names), ORG (organizations, companies, institutions, political groups, and football clubs), and LOC (geographical places), utilizing the IOB2 labeling scheme. In

the provided data, there are 576 instances categorized as "Person" in the training set, with 86 instances in the test set. Additionally, there are 186 "Organization" instances in the training set and 20 in the test set. Finally, the training set contains 548 instances labeled as "Location" while the test set has 113 such instances. In total, there are 1310 entities in the training set and 219 entities in the test set.

39.3.2 Hyper-Parameters Setting

Our models are implemented using PyTorch [16] and Python 3.7. The training process integrates the backpropagation algorithm to iteratively update the parameters of all models. We utilize the Adam optimization algorithm [17] with a learning rate set to 1e-4. Each iteration involves batching the entire training dataset and processing one batch at a time.

In our experiments, we train the model with a batch size of 64 for 30 epochs, incorporating a patience rate of 10. The training and test datasets adhere to the standard split outlined in the original papers, with 80% of the training data allocated for model training and 20% for development and hyperparameter tuning. Model evaluations are conducted using the test dataset.

39.4 Experiments and Results Analysis

In this section, we delve into the experimental outcomes stemming from the application of various architectures incorporating Multi-dialect-Arabic-BERT, Bi-GRU, and CRF on the DzNER-corpus dataset. Our experimental setup encompasses three main phases:

Firstly, we undertake an exhaustive exploration to identify the optimal architecture that yields the highest performance. This involves testing different combinations and configurations of the aforementioned components to discern the most effective arrangement for the NER task in the context of the Algerian Arabic dialect. Subsequently, we meticulously analyze and discuss the outcomes obtained from our experiments. Finally, we delve into a comprehensive error analysis to gain deeper insights into the model's performance and identify areas for potential improvement.

39.4.1 Result

To assess the performance and evaluate the impact of different architectural configurations, we conducted a comprehensive series of comparative experiments. Specifically, we explored four distinct architectures:

1. Fine-tuning the Multi-dialect-Arabic-BERT model.
2. Concatenating Multi-dialect-Arabic-BERT with Bi-GRU.
3. Incorporating Multi-dialect-Arabic-BERT with CRF.
4. Integrating Multi-dialect-Arabic-BERT, Bi-GRU, and CRF layers.

In our comparative analysis, we assessed the performance of each architecture based on well-established metrics including precision, recall, and F1-score. The results of these experiments are visualized in Figs. 39.2, 39.3, 39.4 and 39.5, where 'P' denotes precision, 'R' signifies recall, and 'F1' indicates the F1-score.

The mentioned figures illustrate the performance metrics of different architectures in the context of NER tasks.

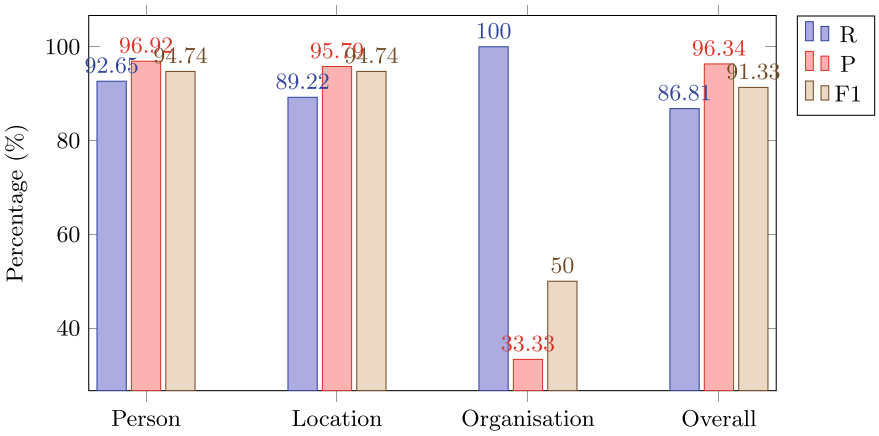


Fig. 39.2 Multi-dialect-Arabic-BERT model performance

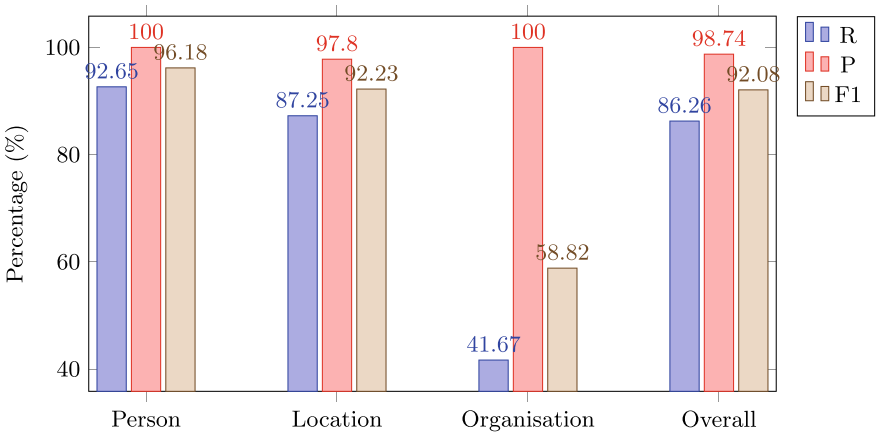


Fig. 39.3 Multi-dialect-Arabic-BERT-Bi-GRU model performance

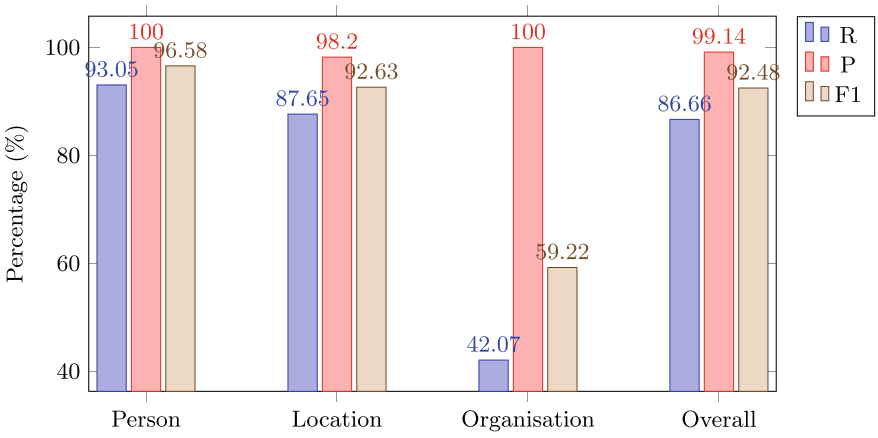


Fig. 39.4 Multi-dialect-Arabic-BERT-CRF model performance

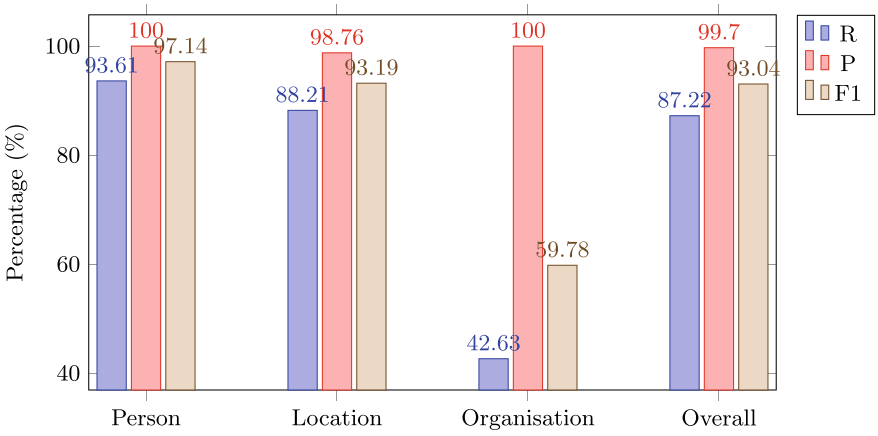


Fig. 39.5 Multi-dialect-Arabic-BERT-Bi-GRU-CRF model performance

Figure 39.2 presents the results for the Multi-dialect-Arabic-BERT model, showcasing varying precision across different entity categories. Notably, the model achieves high recall for organization entities (100%) but demonstrates lower precision (33.33%) and overall (91.33%) F1-score.

Moving to Fig. 39.3, which depicts the Multi-dialect-Arabic-BERT-Bi-GRU model, we observe improvements in precision across all categories compared to the previous model. Particularly, there are notable enhancements in precision for location (97.8%) and overall (92.08%) F1-score, indicating the effectiveness of the Bi-GRU architecture in capturing sequential dependencies.

In Fig. 39.4, we analyze the performance of the Multi-dialect-Arabic-BERT-CRF model. This architecture demonstrates competitive precision across various entity

types, with significant improvements in recognizing geographical locations (92.63% F1-score). The incorporation of the CRF layer enriches the model's ability to decode complex linguistic structures, resulting in enhanced entity recognition.

Finally, Fig. 39.5 outlines the performance metrics of the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model. This architecture combines the benefits of both Bi-GRU and CRF layers, resulting in overall precision improvements across all entity categories. Particularly, the model exhibits high precision for location (98.76%) and overall (93.04%) F1-score, indicating its effectiveness in capturing sequential dependencies and decoding complex linguistic structures.

In summary, while all architectures show promising results, the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model stands out as the most effective in achieving high precision across various entity types. The integration of Bi-GRU and CRF layers enhances the model's ability to capture sequential dependencies and decode complex linguistic structures, ultimately improving entity recognition performance.

39.4.2 Error Analysis

This subsection aims to explore the nature and distribution of errors present in the output generated by our deep learning model. Building upon the superior performance of the Multi-Dialect Arabic-BERT-Bi-GRU-CRF model and inspired by a similar methodology outlined in [18], we analyze errors across five distinct categories: No annotation, No extraction, Wrong range, Wrong tag, and Wrong range and tag.

To elucidate, an output entity from an NER system is considered correct if both its beginning and end positions align with those in the gold data, along with matching tags. Deviation from this criterion constitutes an error. Hence, we classify errors into five types as follows:

- No annotation: This error occurs when the model predicts tokens as a named entity even though those tokens were not annotated in the hand-labeled text.
- No extraction: In this error type, a hand-labeled entity is not predicted by the model.
- Wrong range: This error involves the model incorrectly predicting the boundaries of entities.
- Wrong tag: Here, the model extracts tokens as a named entity with the correct span, but the assigned tag type is incorrect.
- Wrong range and tag: This error denotes cases where the predicted tokens exhibit both incorrect range and tag type assignments.

Table 39.1 illustrates the breakdown of errors by their respective types. These errors were tallied using a logical sum (OR) approach applied to the gold labels and predicted labels. Based on the provided table summarizing errors generated by the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model, several insights can be drawn. Firstly, the table indicates that the most prevalent error type is "Wrong range" accounting for

Table 39.1 Summary of errors

Error type	Number	Rate (%)
No extraction	25	11.68
No annotation	2	0.93
Wrong range	157	73.36
Wrong tag	22	10.28
Wrong range and tag	8	3.73
All errors	214	100

73.36% of all errors. This suggests that the model frequently misjudges the boundaries of named entities. Such a high rate of "Wrong range" errors could stem from the complexity of the text or the nuances within the Algerian Arabic dialect, challenging the model's ability to precisely delineate entity boundaries. Additionally, "No extraction" errors constitute 11.68% of all errors, implying instances where the model fails to identify entities present in the text. This might be due to the diversity of expressions or the presence of uncommon terms not adequately represented in the model's training data. Furthermore, "Wrong tag" errors account for 10.28% of all errors, indicating cases where the model assigns incorrect tags to identified entities. This type of error could result from the model's inability to capture subtle semantic distinctions or context-specific meanings within Algerian Arabic. Lastly, "Wrong range and tag" errors, representing 3.73% of all errors, signify instances where the model misidentifies both the boundaries and the types of named entities. These errors could stem from the intricate syntactic structures or ambiguous linguistic features inherent in Algerian Arabic.

39.5 Conclusion

In conclusion, our study addresses the critical task of NER within the context of Algerian Arabic, a dialect that has received limited attention in NER research despite its linguistic significance. Through the integration of the Multi-dialect BERT model, Bi-GRU, and CRF layer, we propose a novel architecture tailored to the complexities of Algerian Arabic text. Our experimental results highlight the effectiveness of the proposed architectures in accurately identifying named entities within Algerian Arabic text. Among them, the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model stands out with an impressive 93.04% F1-score. This accomplishment underscores the model's capability to capture intricate sequential dependencies and decode the nuanced linguistic structures present in Algerian Arabic. However, despite the robust performance of the Multi-dialect-Arabic-BERT-Bi-GRU-CRF model in NER, our analysis reveals significant challenges in correctly identifying entity boundaries,

assigning accurate tags, and extracting entities from the text. These findings emphasize the necessity for further refinement and adaptation of NER models to effectively address the complexities inherent in Algerian Arabic text.

References

1. Harrat, S., Meftouh, K., Abbas, M., Smaili, K.: Building resources for Algerian Arabic dialects. In: 15th Annual Conference of the International Communication Association Interspeech (2014)
2. Touileb, S.: NERDz: a preliminary dataset of named entities for Algerian. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, pp. 95–101 (2022)
3. Dahou, A.H., Cheragui, M.A.: DzNER: a large Algerian named entity recognition dataset. Natl. Lang. Process. J. **3**, 100005 (2023)
4. Dahou, A.H., Cheragui, M.A.: Named entity recognition for Algerian Arabic dialect in social media. In: International Conference on Computing and Information Technology, pp. 135–145. Springer (2022)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **26** (2013)
6. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
8. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv:1802.05365 (2018)
9. Talafha, B., Ali, M., Za'ter, M.E., Seelawi, H., Tuffaha, I., Samir, M., Farhan, W., Al-Natsheh, H.T.: Multi-dialect Arabic BERT for country-level dialect identification (2020)
10. Goyal, P., Pandey, S., Jain, K.: Deep Learning for Natural Language Processing. Apress, New York (2018)
11. Affi, M., Latiri, C.: Arabic named entity recognition using variant deep neural network architectures and combinatorial feature embedding based on CNN, LSTM and BERT. In: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pp. 302–312 (2022)
12. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv:1603.01354 (2016)
13. Liu, L., Shang, J., Ren, X., Xu, F., Gui, H., Peng, J., Han, J.: Empower sequence labeling with task-aware neural language model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
14. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991 (2015)
15. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
18. Affi, M., Latiri, C.: BE-BLC: BERT-ELMO-based deep neural network architecture for English named entity recognition task. Procedia Comput. Sci. **192**, 168–181 (2021)

Author Index

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies 411, https://doi.org/10.1007/978-981-97-7419-7