Lecture Notes in Networks and Systems 1393

Fausto Pedro García Márquez
Alaa Ali Hameed
Akhtar Jamil *Editors*

Pattern Recognition and Artificial Intelligence

Selected Papers from the 6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI24)

Springer

Lecture Notes in Networks and Systems

Volume 1393

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

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

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

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

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

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

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

Fausto Pedro García Márquez · Alaa Ali Hameed ·
Akhtar Jamil
Editors

Pattern Recognition and Artificial Intelligence

Selected Papers from the 6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI24)

Springer

Editors
Fausto Pedro García Márquez
Department of Business Management
University of Castilla-La Mancha
Ciudad Real, Spain

Alaa Ali Hameed
Department of Computer Engineering
Istinye University
Istanbul, Türkiye

Akhtar Jamil
National University of Computer and Emerging Sciences
Islamabad, Pakistan

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-90892-7 ISBN 978-3-031-90893-4 (eBook)
https://doi.org/10.1007/978-3-031-90893-4

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

Committees

General Chairs

Fausto Pedro García Márquez, Castilla-La Mancha University, Spain
Mehmet Alper Tunga, Istinye University, Türkiye
Alaa Ali Hameed, Istinye University, Türkiye

Steering Committee

Hatice Gülen, Istinye University, Türkiye
Ahmed Abdelgawad, Central Michigan University, USA
Javad Sadri, Concordia University, Canada
Haikal El Abed, German International Cooperation, Lebanon

Program Chairs

Chawki Djeddi, Larbi Tebessi University, Algeria
Akhtar Jamil, National University of Computer and Emerging Sciences, Islamabad, Pakistan
Muhammad Davud, Istinye University, Türkiye

Organizing Committee

Sobhan Sarkar, University of Edinburgh, UK
Wadhah Zeyad Tareq, Istinye University, Türkiye

S. B. Goyal, City University of Malaysia, Malaysia
Parameshachari B. D., Visvesvaraya Technological University, India
Amir Seyyedabbasi, Istinye University, Türkiye

Publication and Publicity Chairs

Rasha Gargees, Central Michigan University, USA
Pasquale Ardimento, Università degli Studi di Bari, Italy

Program Committee

Abbas Cheddad, Blekinge Institute of Technology, Sweden
Abdelhakim Hannousse, Guelma University, Algeria
Abhishek Gupta, University of Washington, USA
Ahmad Bouridane, Northumbria University, UK
Ahmad Din, COMSATS University, Islamabad, Pakistan
Aladine Chetouani, Polytech Orléans, France
Ameur Bensefia, Higher Colleges of Technology, UAE
Amina Pramanik, Indian Institute of Technology, India
Amutha Benziker, Saveetha Engineering College, India
Arif Ur-Rahman, Bahria University, Pakistan
Bart Lamiroy, University of Lorraine, France
Cemal Okan Sakar, Bahçeşehir University, Türkiye
Chayan Halder, West Bengal State University, India
Faezeh Soleimani, Ball State University, USA
Faisal Shafait, National University of Sciences and Technology, Pakistan
Gabriella Casalino, Università Degli Studi Di Bari, Italy
Gattal Abdeljalil, Larbi Tebessi University, Algeria
Haikal El Abed, Technical Trainers College, German International Cooperation, Lebanon
Hamid Laga, Murdoch University, Australia
Hammad Afzal, National University of Sciences and Technology, Pakistan
Hassiba Nemmour, University of Science and Technology Houari Boumediene, Algeria
Hedi Tabia, University of Paris Saclay, France
Imran Siddiqi, Bahria University, Pakistan
Jean-Marc Ogier, University of la Rochelle, France
Jocelyn Chanussot, Grenoble Institute of Technology, France
Juan Humberto Sossa Azuela, Instituto Politécnico Nacional, Mexico

Khadim Kamil Resan, Al-Mustansirya University, Iraq
Khurram Khurshid, Institute of Space Technology, Pakistan
Kiruthika Jayaramakrishnan, Kingston University, UK
Kiran Sood, Chitkara University, India
Laurent Heutte, University of Rouen, France
Lincoln Silva, Universidade Federal Fluminense, Brazil
Lotfi Chaari, University of Toulouse, France
Marco Anisetti, Università degli Studi di Milan, Italy
Osama M. F. Dawoud, Istinye University, Türkiye
Mohamed Chaouki Babahenini, Biskra University, Algeria
Mohamed Cherif Nait Hamoud, Larbi Tebessi University, Algeria
Mohamed El Bachir Menai, King Saud University, KSA
Moises Diaz, Universidad de Las Palmas de Gran Canaria, Spain
Momina Moetesum, Bahria University, Pakistan
Muhammad Abdul Basit, Montana Technological University, Butte Montana, USA
Muhammad Asfand-E-Yar, Bahria University, Pakistan
Muhammad Ilyas, Al Ain University, Abu Dhabi, UAE
Muhammad Imran Malik, National University of Sciences and Technology, Pakistan
Muhammad Shehzad Hanif, King Abdulaziz University, KSA
Muhammed Cinsdikici, Manisa Celal Bayar University, Türkiye
Murat Gezer, Istanbul University, Türkiye
Mustafa Ali Abuzaraida, University of Utara Malaysia, Malaysia
Mustafa Takaoğlu, TÜBİTAK-BİLGEM, Kocaeli, Türkiye
Mustansar Ali Ghazanfar, University of East London, UK
Nicole Vincent, Paris Descartes University, France
Nelly Elsayed, University of Cincinnati, USA
Neha Sharma, Chitkara University, India
Okba Kazar, University of Biskra, Algeria
Paulo Batista, University of Évora, Portugal
Pinar Kirci, Uludag University, Türkiye
Renjith V. Ravi, MEA Engineering College, Kerala, India
Said Ghoul, Philadelphia University, Jordan
Şebnem Özdemir, Istanbul İstinye Univerity, Istanbul, Türkiye
Sergey Ablameyko, Belarusian State University, Belarus
Shadi Alzubi, Al-Zaytoonah University of Jordan, Jordan
Shuhaida Shuhidan, Universiti Teknologi, Malaysia
Sourour Ammar, Digital Research Center of Sfax, Tunisia
Subhan Ullah, National University of Computer and Emerging Sciences, Pakistan
Sujata Patil, Mumbai University, India
Thierry Paquet, University of Rouen, France
Vaibhav Lodhi, IIT Kharagpur, India
Vinayakumar Ravi, Prince Mohammad Bin Fahd University, KSA
Vincent Christlein, University of Erlangen-Nuremberg, Germany

Virendra Kuma, All India Institute of Medical Sciences, India
Waleed Ead, Beni Suef University, Egypt
Yan Chai Hum, Universiti Tunku Abdul Rahman, Malaysia
Yassine Ben Ayed, ISIMS Sfax, Tunisia
Youcef Chibani, University of Science and Technology Houari Boumediene, Algeria

Preface

This is a Proceedings book which will include the best papers of the 6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI24), Istinye University, Istanbul, Türkiye, on October 18–19, 2024. The book is unique due to the presence of some state-of-the-art contributions related to emerging topics and applications in the domain of artificial intelligence. Some recent advancements in these fields are included in this book.

It is with great pleasure and anticipation that we extend our warm welcome to you all for the 6th edition of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI24), scheduled to take place at.

Building on the success of our previous conferences in Algeria, Morocco, Tunisia, and Türkiye, MedPRAI has evolved into a premier platform for researchers and practitioners from academia and industry to converge and exchange ideas on the latest developments in pattern recognition and artificial intelligence. The conference is proudly organized by the Istinye University, Istanbul, Türkiye, and our mission is to foster an environment conducive to insightful discussions and collaborative initiatives.

To ensure the high quality of the conference, all submitted papers were undergo a thorough review process by at least three members of the international program committee, with additional reviewers consulted if necessary. The manuscripts were evaluated based on criteria such as originality, significance, technical soundness, and clarity of presentation.

MedPRAI offered a forum for researchers proposing solutions meeting the operational needs of industry to meet the industry's operational needs. The conference was hence targeted both at researchers working on innovative technology and experts developing tools in the field. The objective of the conference was to attract papers investigating the use of technology outside the controlled environment of research laboratories. It intended to offer a venue for theory and data-driven techniques.

The conference facilitated the young researchers, industries, and research agencies especially, who were carrying out their research work in the domain of Computer Science, Information Technology, Electrical, Electronics and Communication Engineering with valuable discussions in order, those who were carrying out their research work in the domain of Computer Science, Information Technology, Electrical, Electronics and Communication Engineering with valuable discussions to make the outcomes more realistic.

The topics are grouped into three main thematic areas. Topics of interest scope all related subareas including but not limited to the following:

1. **Artificial Intelligence and Applications**
   - Machine Learning and Data Mining
   - Natural Language Processing
   - Decision Support Systems
   - Intelligent Automation in Manufacturing
   - AI in Healthcare
   - Financial Forecasting with AI and Predictive Analytics
   - Business Intelligence Enhanced by AI
   - AI-powered Customer Relationship Management (CRM)
   - Artificial Intelligence in Teaching Methods and Education
   - Artificial Intelligence in Chemistry Sciences and Material Discovery
   - Artificial Intelligence in Media Studies and Historical Studies
   - Character, text and graphics recognition
   - Document Analysis and Processing
   - Artificial Intelligence in Linguistics, Literary, and Religious Analysis
   - Artificial Intelligence in Sports Training and Exercise
   - AI for Personalized Learning and Curriculum Development
   - AI in Environmental Science and Sustainability
   - AI for Predictive Analytics and Decision-Making in Business
   - AI in Sports Analytics and Performance Enhancement
   - Artificial Intelligence in Economics, Accounting and Finance
2. **Computer Vision and Image Processing**
   - Self-Supervised Learning
   - Transformer Architectures
   - Image Synthesis and Augmentation
   - Explainable AI in Vision
   - Edge Computing in Vision Systems
   - Zero-shot Learning
   - Semantic Segmentation with Deep Learning
   - Transfer Learning in Computer Vision
   - Multi-modal Learning in Computer Vision
   - Image analysis, coding and processing
   - Industrial Applications of Pattern Recognition
   - 3D Vision and Geometry
   - Computer Vision for Autonomous Vehicles
   - Deep Learning for Video Compression
   - Real-time Object Detection and Tracking
   - AI for Drone-based Imaging

- Neural Networks for Document Image Analysis
- Augmented Reality and Virtual Reality Applications
- Facial Recognition Technology and its Ethical Implications
- Deep Fakes Detection and Prevention Techniques

3. **Blockchain, IoT, and Internet of Things**

- Decentralized Finance (DeFi) with Blockchain
- Next-gen Decentralized Applications (DApps) Architectures
- Smart Contracts 2.0: Beyond Traditional Use Cases
- Interoperability Challenges in Blockchain-IoT Integration
- Privacy-Preserving Approaches in IoT Networks
- AI-Enabled Industrial IoT (IIoT) Innovations
- Sustainable Supply Chains with Blockchain Solutions
- Scalability Solutions for Blockchain Networks
- Security and Privacy in IoT and Blockchain
- Edge Computing Integration in IoT and Blockchain
- Tokenization of Physical Assets on Blockchain
- Blockchain for Healthcare: Applications in Medical Records and Pharmaceuticals
- IoT in Agriculture: Smart Farming Solutions
- IoT Device Authentication and Authorization
- Quantum-Resistant Blockchain Technologies
- Blockchain in Government and Public Services
- IoT in Smart Cities: Infrastructure and Services
- Energy-Efficient Protocols for IoT Devices
- Blockchain for Voting Systems: Transparency and Security
- IoT and Blockchain for Disaster Management and Response

Ciudad Real, Spain — Prof. Fausto Pedro García Márquez
Islamabad, Pakistan — Dr. Akhtar Jamil
Istanbul, Türkiye — Dr. Alaa Ali Hameed

Contents

About the Editors

Prof. Fausto Pedro García Márquez works at UCLM as a **Full Professor** (Accredited as Full Professor from 2013), Spain, an **Honorary Senior Research Fellow at Birmingham University**, UK, a **Lecturer at the Postgraduate European Institute**, a Research Fellow at **INTI International University and Colleges**, Malaysia, and has been **Senior Manager in Accenture** (2013–2014). He obtained his European Ph.D. with maximum distinction. He has been **distinguished with the prices**: Runner Prize (2023), Nominate Prize (2022), Gran Maestre (2022), Grand Prize (2021), Runner Prize (2020) and Advancement Prize (2018), Runner (2015), Advancement (2013) and Silver (2012) by the International Society of Management Science and Engineering Management (ICMSEM), First International Business Ideas Competition 2017 Award (2017), etc. **He has published more than 294 journal papers** (202 JCR: 102-Q1; 52-Q2; 39-Q3; 9-Q4), some recognized as: *Progress in Photovoltaics: Research and Applications* (Q1, IF. 8.49, one of the most downloaded in first 12 months of publications, 2023), *Applied Energy* (Q1, IF 9.746, as "Best Paper 2020"), *Renewable Energy* (Q1, IF 8.001, as "Best Paper 2014"); *ICMSEM* (as "excellent"), *International Journal of Automation and Computing* and *IMechE Part F: Journal of Rail and Rapid Transit* (most downloaded), etc. He is the author and editor of over **62 books** (Elsevier, Springer, Pearson, Mc-GrawHill, Intech, IGI, Marcombo, AlfaOmega), > 115 international chapters, and 6 patents. He is the Editor of 5 International Journals, Committee Member more than 70 International Conferences. He has been the **Principal Investigator in 5 European Projects, 8 National Projects, and more than 150 projects for Universities, Companies, etc**. He is being: Expert in the European Union in AI4People (EISMD), and ESF; **Director of** https://ingenium.uclm.es/**.** Senior Member at IEEE, 2021-… Honoured Honourary Member of the Research Council of Indian Institute of Finance, 2021-… Committee Chair of The International Society for Management Science and Engineering Management (ISMSEM), 2020-… Research Fellow at **INTI International University and Colleges**, Malaysia, from 2023-… Visiting Professor at **Sheyang University (China)** from 2024 to 2027. Advisory Board of the EU Project SunBio (Id: ID 101157493) 2024–2027. His main interests are: Artificial Intelligence, Maintenance, Management, Renewable Energy, Transport, Advanced Analytics, and Data Science.

Dr. Alaa Ali Hameed obtained his master's degree in Computer Engineering from Eastern Mediterranean University in North Cyprus, followed by a Ph.D. from Selcuk University in Türkiye. He completed his Postdoctoral research at Istanbul University. He has held academic positions as an Assistant Professor at Istanbul Aydin University and Istanbul Sabahattin Zaim University and currently serves in the same capacity at Istinye University. Hameed has an impressive record of over 100 publications in prestigious international journals and conferences, demonstrating his prolific contribution to the field. He has also taken on significant leadership roles, including serving as Conference Chair, Program Chair, and Technical Program Committee member for various international conferences. In addition, he has contributed to the academic community as an editor for several books and as a reviewer for numerous journals. His editorial contributions extend to serving as a Guest Editor for several SCIE-indexed journals. Hameed's research interests are diverse and include Data Mining, Machine/Deep Learning, Big Data Analytics, Artificial Intelligence, Digital Signal/Image Processing, and Adaptive Filters. He has actively participated in numerous research activities, further establishing his expertise and influence in these areas.

Dr. Akhtar Jamil is an accomplished Associate Professor at the National University of Computer and Emerging Sciences in Islamabad, Pakistan. He holds a Ph.D. in Computer Science from Yildiz Technical University, Istanbul, and a Postdoctoral Fellowship from Istanbul University-Cerrahpasa, Türkiye. Dr. Jamil has an extensive work history, including key academic and industry roles. Before his current position, he served as an Assistant Professor and Deputy Head of the Department of Computer Engineering at Istanbul Sabahattin Zaim University, Türkiye. He also held positions in international companies such as EMI Group Information Technology in Türkiye and collaborated with the Ingenium Research Group at the University of Castilla-La Mancha, Spain, as a Visiting Professor.

Dr. Jamil is an active researcher; he has authored 28 journal papers, 8 book chapters, and 71 conference papers, making significant contributions to his field. He has been actively involved in leading international research collaborations and has played key roles in organizing major conferences as Technical Chair and General Chair. He is also an editor of several academic books and a regular reviewer for prestigious journals, further solidifying his impact on the global computer science community. His current research interests focus on Artificial Intelligence, Computer Vision, Generative AI, and Pattern Recognition.

Comparative Analysis of Machine Learning Algorithms for Phishing Detection

Dimitar Rangelov, Radoslav Miltchev, Evgeni Genchev, and Petar Kirkov

Abstract Phishing attacks represent a significant threat in the realm of cybercrime, characterized by the use of deceptive emails and websites to extract sensitive information from unsuspecting individuals. Traditional detection methods have become less effective against increasingly sophisticated phishing techniques, necessitating the adoption of advanced technologies such as machine learning (ML) for enhanced detection capabilities. This research paper presents a comparative analysis of various machine learning algorithms—Naive Bayes, Logistic Regression, SGD Classifier, XGBoost, Decision Tree, Random Forest, and MLPClassifier—focusing on their effectiveness in phishing email classification. By evaluating key performance metrics such as accuracy, validation accuracy, classification reports, and confusion matrixes, the study identifies the strengths and weaknesses of each algorithm. The results highlight the SGD Classifier and Logistic Regression as the most robust models, exhibiting high generalization capabilities suitable for practical applications. Conversely, models like XGBoost and Decision Tree showed possible overfitting, indicating a need for further tuning. The findings provide valuable insights for cybersecurity professionals, aiding in the selection and implementation of effective phishing detection systems. Future research should explore hybrid models and

D. Rangelov (✉)
AI and CAD Systems Lab at R & D & I Consortium (Sofia Tech Park), University of Twente, Enschede, The Netherlands
e-mail: d.g.rangelov@utwente.nl

R. Miltchev
AI and CAD Systems Lab at R & D & I Consortium (Sofia Tech Park), Technical University of Sofia, Sofia, Bulgaria
e-mail: rmiltchev@tu-sofia.bg

E. Genchev
AI and CAD Systems Lab at R & D & I Consortium (Sofia Tech Park), Sofia, Bulgaria
e-mail: e.a.genchev@saxion.nl

P. Kirkov
Technical University of Sofia, Sofia, Bulgaria
e-mail: pkirkov@gmail.com

F. P. García Márquez et al. (eds.), *Pattern Recognition and Artificial Intelligence*, Lecture Notes in Networks and Systems 1393,
https://doi.org/10.1007/978-3-031-90893-4_1

advanced feature extraction techniques to further enhance detection accuracy and robustness.

Keywords Phishing detection · Machine learning · Email classification

1 Introduction

Phishing attacks have become one amongst the most pervasive and malicious cybercrimes over the past couple of years. It is a type of attack where deceptive emails, websites, or messages are targeted towards people in order to extract sensitive information like usernames, passwords, or credit card numbers. The ease of conducting an attack and the potential for high returns make phishing a very significant threat to both individuals and organizations. Studies show that a high percentage of data breaches and financial losses are being attributed to phishing attacks, hence the pressing need to develop strong detection mechanisms aimed at minimizing the same.

Conventional ways have been losing their effectiveness with the increasing sophistication of phishing. Attackers always keep on coming up with new strategies to bypass security measures. Thus, advanced technologies should be brought in detecting phishing. Machine learning has emerged to great effect in this respect as a strong tool that can easily handle huge volumes of data and identify patterns that might be indicative of phishing. Despite the potential of ML in enhancing phishing detection, there is a need for a comprehensive comparison of various ML algorithms to determine their relative effectiveness. Understanding the strengths and weaknesses of different algorithms can inform the development of more robust and accurate phishing detection systems.

The primary objective of this research paper is to perform a comparative analysis of various machine learning algorithms for phishing detection. By evaluating the performance of several widely used algorithms, including Naive Bayes, Logistic Regression, SGD Classifier, XGBoost, Decision Tree, Random Forest, and MLPClassifier, this study aims to identify which methods are most effective in distinguishing phishing emails from legitimate ones.

This study seeks to address the following research questions:

1. How do different machine learning algorithms compare in terms of accuracy for phishing detection?
2. What are the common strengths and weaknesses of the evaluated machine learning models in the context of phishing detection?
3. How can the findings of this comparative analysis inform the development of more effective phishing detection systems?

This research paper makes significant contributions both theoretically and practically. Theoretically, it advances the understanding of the efficacy of various machine learning algorithms in the context of phishing detection. By providing a detailed comparative analysis, it adds to the existing body of knowledge on the strengths and

Table 12 Scenario B: DKASC A dataset results for 5%, 10% and 15% missing data fractions

Method	5%			10%			15%		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MHD	1.886	2.124	0.213	1.935	2.185	0.187	1.946	2.198	0.191
LVCF	0.548	1.067	0.801	0.580	1.082	0.801	0.630	1.146	0.780
LI	0.394	0.789	0.891	0.424	0.801	0.891	0.453	0.859	0.876
KNN	0.279	0.480	0.960	0.312	0.521	0.954	0.327	0.547	0.950
SVM	0.532	0.622	0.933	0.532	0.627	0.933	0.535	0.645	0.930
XGBoost	0.226	0.397	0.973	0.252	0.425	0.969	0.267	0.443	0.967
LGBM	0.237	0.409	0.971	0.253	0.424	0.969	0.268	0.445	0.967
RF	0.226	0.413	0.970	0.234	0.412	0.971	0.246	0.432	0.969
ERTR	**0.210**	**0.381**	**0.975**	**0.228**	**0.405**	**0.972**	**0.241**	**0.423**	**0.970**
MICE	0.726	1.009	0.823	0.824	1.102	0.793	0.923	1.203	0.757

Table 13 Scenario B: DKASC B dataset results for 5%, 10% and 15% missing data fractions

Method	5%			10%			15%		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MHD	4.717	5.571	0.268	4.759	5.596	0.257	4.839	5.688	0.236
LVCF	1.356	2.485	0.854	1.370	2.472	0.855	1.435	2.547	0.847
LI	0.954	1.826	0.921	0.969	1.817	0.922	1.011	1.882	0.916
KNN	0.803	1.312	0.959	0.849	1.356	0.956	0.866	1.401	0.954
SVM	1.163	1.436	0.951	1.185	1.463	0.949	1.191	1.496	0.947
XGBoost	0.688	1.129	0.970	0.706	1.134	0.969	0.745	1.207	0.966
LGBM	0.689	1.134	0.970	0.709	1.148	0.969	0.726	1.181	0.967
RF	0.635	1.069	0.973	0.667	1.121	0.970	0.686	1.146	0.969
ERTR	**0.642**	**1.066**	**0.973**	**0.661**	**1.102**	**0.971**	**0.679**	**1.160**	**0.968**
MICE	1.692	2.527	0.849	2.011	2.844	0.808	2.288	3.129	0.769

over time and LI assuming linear transitions between points—assumptions that may not hold in datasets with trends or seasonal effects. These methods can introduce bias; for instance, MHD's mean replacement might skew data distributions, and LVCF assumes missing values are identical to the previous ones, which is unreliable in dynamic datasets. Additionally, LVCF and LI over-rely on temporal proximity, making them susceptible to inaccuracies if adjacent data points are anomalies or if there are large gaps.

Table 14 Scenario B: ASHRAE B-276 results for 5%, 10% and 15% missing data fractions

Method	5%			10%			15%		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MHD	27.979	33.424	0.166	25.495	30.495	0.269	25.569	30.326	0.344
LVCF	7.822	11.051	0.909	8.414	14.885	0.826	8.315	14.040	0.859
LI	3.612	4.751	0.983	4.022	6.996	0.962	4.004	6.485	0.970
KNN	5.212	7.732	0.955	5.783	8.471	0.944	7.157	10.372	0.923
SVM	11.429	14.394	0.845	12.058	15.081	0.821	13.436	16.448	0.807
XGBoost	4.014	6.391	0.970	4.830	10.466	0.914	4.829	8.627	0.947
LGBM	3.730	5.100	0.981	4.713	10.214	0.918	4.449	7.949	0.955
RF	3.445	4.829	0.983	3.897	6.677	0.965	4.150	6.602	0.969
ERTR	**3.348**	**4.428**	**0.985**	**3.533**	**5.586**	**0.975**	**3.900**	**5.616**	**0.977**
MICE	9.242	15.286	0.826	10.704	20.219	0.679	12.243	21.729	0.663

Table 15 Scenario B: ASHRAE B-505 results for 5%, 10% and 15% missing data fractions

Method	**5%**			10%			15%		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MHD	7.679	9.340	0.141	8.099	9.679	0.041	7.934	9.532	0.090
LVCF	2.349	4.444	0.806	3.084	5.858	0.649	3.163	5.869	0.655
LI	2.207	3.343	0.890	2.305	3.569	0.870	2.548	3.903	0.847
KNN	1.512	2.152	0.954	1.851	2.871	0.916	2.237	3.573	0.872
SVM	2.669	3.353	0.889	2.864	3.666	0.862	3.298	4.398	0.806
XGBoost	1.307	2.052	0.959	1.469	2.252	0.948	1.727	2.801	0.921
LGBM	**1.243**	**1.670**	**0.973**	**1.434**	**2.317**	**0.945**	**1.737**	**2.639**	**0.930**
RF	1.252	1.811	0.968	1.472	2.349	0.944	1.688	2.751	0.924
ERTR	1.234	1.829	0.967	1.492	2.425	0.940	1.729	2.975	0.911
MICE	2.609	4.169	0.829	3.563	6.143	0.614	4.204	7.382	0.454

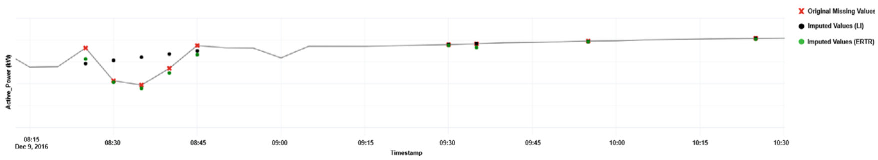

Fig. 2 ERTR versus LI imputation performances on a chunk of DKASC A dataset

7 Conclusion

In this study, we introduced a methodology for missing data imputation and conducted an empirical comparison of seven imputation methods on two types of incomplete datasets: photovoltaic energy production and energy consumption of different types of buildings.

Our findings revealed several key insights. First, the inclusion of lagged variables in machine learning-based imputation methods significantly improves performance. However, we observed diminishing returns with higher lags, indicating that beyond a certain point, increasing the number of lagged variables results in progressively smaller performance gains, especially when missing values are adjacent to two existing values. In scenarios where there are consecutive blocks of missing values and the lagged variables also have consecutive missing values, the effectiveness of introducing lagged variables diminishes.

Our study also found that bagging methods, specifically Extremely Randomized Trees (ERTR), consistently outperformed other ML methods. In contrast, linear interpolation (LI) was the best-performing statistical-based imputation method.

Our study also confirmed that as the fraction of missing values in the dataset increases, achieving high imputation performance becomes more challenging.

Finally, our results indicate that machine learning-based imputation methods outperform statistical-based techniques, particularly in datasets that include sudden spikes and plummets.

The study is, of course, subject to limitations and provides room for further research. First, we will add more features that might impact energy imputation such as buildings orientation, building shape, passive heating, and cooling mechanisms. Second, as ML models evolve and new models are introduced, we aim to evaluate the performances of deep generative models such as Variational Autoencoder and Generative Adversarial Networks models in addition to their combination.

References

1. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021) A survey on missing data in machine learning. J Big Data 8:1–37
2. Rubin D-B (1976) Inference and missing data. Biometrika 63(3):581–592
3. Ren L, Wang T, Seklouli A-S, Zhang H, Bouras A (2023) A review on missing values for main challenges and methods. Inf Syst 119:102268
4. Tian C, Ma J, Zhang C, Zhan P (2018) A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network. Energies 11(12):3493
5. Haddad H, Jerbi F, Smaali I (2024) Toward unsupervised energy consumption anomaly detection. In: International conference on artificial intelligence applications and innovations. Springer Nature, Switzerland, Greece, pp 332–346
6. Chong A, Lam K-P, Xu W, Karaguzel O-T, Mo Y (2016) Imputation of missing values in building sensor data. In: ASHRAE & IBPSA-USA SimBuild 2016: building performance modeling conference. ASHRAE/IBPSA-USA, Salt Lake City, pp 407–4014

7. Panapakidis I-P, Aggelos S-B, Georgios C-C (2018) A missing data treatment method for photovoltaic installations. In: 5th IEEE international energy conference. IEEE, Cyprus, pp 1–6
8. Dhungana H, Bellotti F, Berta R, De Gloria A (2021) Performance comparison of imputation methods in building energy data sets. In: Applications in electronics pervading industry. Environment and society. Springer International Publishing, Hybrid, pp 144–151
9. Tsai C-F, Hu Y-H (2022) Empirical comparison of supervised learning techniques for missing value imputation. Knowl Inf Syst 64(4):1047–1075
10. Batista G-E, Monard M-C (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5–6):519–533
11. Rubin D-B (1976) Inference and missing data. Biometrika 63(3):519–533
12. Krivohlava Z, Chren S, Rossi B (2022) Failure and fault classification for smart grids. Energy Inf 5(1):33
13. Khattab AAR, Elshennawy NM, Fahmy M (2023) GMA: gap imputing algorithm for time series missing values. J Electr Syst Inf Technol 10(1):41
14. Kontopoulou V-I, Panagopoulos A-D, Kakkos I, Matsopoulos G-K (2023) A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Future Internet 15(18):100121
15. Livera A, Theristis M, Koumpli E, Theocharides S, Makrides G, Sutterlueti J, Stein J-S, Georghiou G-E (2021) Data processing and quality verification for improved photovoltaic performance and reliability analytics. Progr Photovolt Res Appl 29(2):143–158
16. Lotfipoor A, Patidar S, Jenkins D-P (2023) Transformer network for data imputation in electricity demand data. Energy Build 300:113675
17. Haddad H, Jerbi F, Smaali I (2024) Learning-based short-term energy consumption forecasting. In: International conference on artificial intelligence applications and innovations. Springer Nature Switzerland, Corfu, Greece, pp 238–251
18. Yuan Y, Zhang C, Wang Y, Liu C, Ji J, Feng C (2017) Linear interpolation process and its influence on the secondary equipment in substations. In: China international electrical and energy conference (CIEEC) (SCDS). IEEE, China, pp 205–209
19. Rahman H-A-A, Wah Y-B, He H, Bulgiba A (2015) Comparisons of ADABOOST, KNN, SVM and logistic regression in classification of imbalanced dataset. In: 1st international conference on soft computing in data science (SCDS). Springer, Putrajaya, Malaysia, pp 54–64
20. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: 22nd ACM SIGKDD international conference on knowledge discovery and data mining. IEEE, CA, USA, pp 785–794
21. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. In: 31st international conference on neural information processing systems. Curran Associates Inc., CA, USA, pp 3149—3157
22. Biau G, Scornet E (2016) A random forest guided tour. Test 25:197–227
23. John V, Liu Z, Guo C, Mita S, Kidono K (2016) XGBoost: real-time lane estimation using deep features and extra trees regression. In: 7th Pacific-Rim symposium (PSIVT). Springer International Publishing, New Zealand, pp 721–733
24. Van Buuren S, Groothuis-Oudshoorn K (2011) Multivariate imputation by chained equations. J Stat Softw 45(3):1–67
25. Botchkarev A (2018) Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. arXiv preprint arXiv:1809.03006. Last accessed 12 Dec 2023
26. DKASC Alice Springs dataset and Yulara Solar System dataset. https://dkasolarcentre.com.au. Last accessed 4 Dec 2024
27. Miller C, Arjunan P, Kathirgamanathan A, Fu C, Roth J, Park J-Y, Balbach C, Gowri K, Nagy Z, Fontanini A-D, Haberl J (2020) The ASHRAE great energy predictor III competition: overview and results. Sci Technol Built Environ 26(10):1427–1447