AI, Big Data, and Decision Sciences Ethical Frameworks, Security, and Inclusive Technologies

Sciforce

ISBN: 979-8-218-59250-9

www.sciforce.org

Sciforce Publications
www.sciforce.org

Ethical Frameworks and Privacy Protection in AI- Driven Big Data: A Comprehensive Analysis Using Support Vector Regression and Ada Boost Models

Authors & Editors
Mr. Sudhakara Reddy Peram.

Correspondence
Engineering Leader, Illumio Inc., United States

Published By
Sciforce Publications
May 12, 2025

Ethical Frameworks and Privacy Protection in AI-Driven Big Data: A Comprehensive Analysis Using Support Vector Regression and Ada Boost Models

Abstract: This study explores the critical intersection of ethics and privacy in AI-driven big data analytics, addressing growing concerns around responsible AI implementation. As AI systems are increasingly integrated into diverse sectors including healthcare, education, and finance, the need for comprehensive ethical frameworks and privacy protection mechanisms is paramount. The research analyzes four key metrics—data sensitivity score, privacy risk index, AI bias metric, and ethical compliance score—over 500 observations to assess the ethical landscape of AI applications. Using advanced machine learning techniques including support vector regression (SVR) and Ada boost regression (ABR), the study demonstrates strong predictive capabilities for the ethical compliance score, with R^2 values of 0.95 and 0.89, respectively. The analysis reveals a significant positive correlation between data sensitivity and ethical compliance (0.76), while AI bias shows a moderate correlation with ethical standards (0.54). Privacy risk demonstrates weak correlations with other variables, suggesting complex relationships in privacy governance. The findings indicate that organizations handling sensitive data tend to maintain high ethical standards, while addressing algorithmic biases contributes to overall ethical AI implementation. The study emphasizes the urgent need for integrated ethical and technical safeguards in AI systems, particularly in contexts that process personal, financial, or health information. These results provide valuable insights for policymakers, AI developers, and organizations seeking to balance technological innovation with ethical responsibility and privacy protection.

Keywords: AI Ethics, Data Privacy, Algorithmic Bias, Ethical Compliance, Big Data Analytics, Machine Learning Governance.

INTRODUCTION

In artificial intelligence, AI ethics involves a formal framework of moral principles and practices that aim to guide the development and use of AI technologies. As AI becomes more deeply integrated a variety of products and services, a growing number of organizations are working to establish dedicated ethical codes. These codes, called AI value platforms, serve as comprehensive policy frameworks that define how artificial intelligence should be used to support progress and improve human well-being [1]. [2] With the advent of technologies like artificial intelligence, big data analytics, personalized medicine continue to evolve, the need for comprehensive education and training among healthcare professionals is increasing. To effectively integrate these innovations into clinical practice, healthcare providers must develop the skills necessary to understand and manage the issues associated with these technologies. [3] Artificial Intelligence (AI) is reshaping the education landscape, fundamentally changing traditional teaching practices, assessment methods, and administrative functions. Tools such as smart learning systems, automated assessment technologies, personalized learning platforms, and predictive data analytics are opening up new avenues for improving educational outcomes.

By enabling personalized learning experiences, AI tailor's instruction to meet the unique needs of each student, thereby increasing engagement and academic performance. [4] It explores the challenges associated with the collection, storage, and analysis of large-scale datasets, while also exploring important concerns related Reducing bias and ensuring transparency in the design of AI algorithms. This in-depth study highlights the critical role of ethical frameworks in shaping the responsible advancement of data-driven AI innovations. [5] The hypothesis put forward in this study asserts that while data-centric AI offers significant opportunities in a variety of fields, it also introduces significant ethical challenges that should not be overlooked. These challenges include important Challenges include protecting privacy, maintaining data integrity, addressing systemic biases, and ensuring transparency in AI-based decision-making processes. [6] The use of artificial intelligence (AI) also raises significant ethical and privacy concerns have sparked growing interest in various sectors of society. For example, the public has expressed fears about the increasing use of robots and their potential impact on unemployment and social inequality. Social scientists have highlighted serious privacy issues, particularly in the context of surveillance systems, while the limited regulation of social media platforms has sparked debates about the misuse of personal data by large technology companies. [7] Predictive policing emerged as a technological solution that aims to prevent crime by predicting potential threats using data-driven analytics. While it represents a modern and forward-looking strategy for law enforcement, its effectiveness relies heavily on collecting and processing extensive datasets, often containing personal and sensitive information. [8]

These regulations go beyond enforcing transparency in data handling; they also provide individuals with enhanced control over their personal information. As a result, organizations should incorporate these legal obligations into their data management policies to reduce the risk of legal consequences. Within this framework, it is crucial to foster a strong sense of ethical responsibility among employees. [9] Existing literature highlights the urgent importance of addressing ethical and legal issues in managing AI-driven IT projects. Challenges such as AI biases require the adoption of comprehensive Techniques such as algorithmic audits and machine learning focused on fairness are increasingly being used. Data privacy continues to be a major concern, forcing organizations to comply with regulations such as GDPR and CCPA, while also implementing AI models that protect user privacy. [10] [11] The Chapter, titled "Navigating the Intersection of Ethics and Privacy in the AI Era," explores the complex relationship between ethical considerations and privacy in the context of artificial intelligence. As AI systems become more pervasive, challenges surrounding data privacy and ethical implications have intensified. The chapter emphasizes the ethical issues involved in handling sensitive data and underscores the need to establish frameworks protect individual rights while fostering technological progress. [12] The chapter, titled "Ethical Considerations and Data Privacy in Artificial Intelligence," explores the complex interplay between privacy and ethics within AI technologies. As AI becomes increasingly pervasive, concerns about data privacy and ethical implications have increased. The chapter highlights the ethical dilemmas

associated with the collection and use of sensitive data, emphasizing the importance of creating frameworks that protect individual rights while fostering technological progress. [13] This comprehensive review examines how artificial intelligence (AI) is contributing to the transformation of big data in financial environments, with a particular focus on the protection of personally identifiable information (PII). Through an analysis of 37 scholarly articles, the study review highlights both the advances and challenges posed by AI technologies in protecting PII during data transformation. [14] The rapid advancement of artificial intelligence (AI) technologies has raised pressing ethical and privacy issues that are particularly relevant in the modern digital era. As AI applications become increasingly embedded in various areas of daily life the collection and analysis of sensitive data raise moral dilemmas that challenge our understanding of privacy and individual rights. [15] The rapid development of artificial intelligence (AI) technology has raised significant ethical and privacy issues, and these issues are becoming increasingly important in the current digital environment. As AI systems are integrated into more aspects of everyday life, the collection and analysis of sensitive data increase moral dilemmas that challenge our understanding of privacy and individual rights. [16] Fintech's artificial intelligence (AI) and big data integration have transformed customer experiences, but it also raises questions about data collection, use, and security.

The paper emphasizes the need to scrutinize these technologies to ensure they do not perpetuate bias or discrimination against certain groups of customers. [17] Big data is a term used to describe the enormous amount of information generated and collected from various sources, made possible by technological advances. It encompasses advances in computational speed, storage capacity, and cost-effectiveness of data collection techniques. The ability to analyze this data can lead to new insights and knowledge that were previously unattainable. [18] Privacy issues are particularly pronounced in this context, as the large amounts of data can lead to identification and inference of sensitive information, challenging traditional privacy protections. Furthermore, the lack of transparency in AI algorithms make it more difficult for people to understand and manage the use of personal data. [19] The emergence of big data and AI in fin tech: The introduction highlights the rapid adoption of artificial intelligence (AI) and big data analytics by the fin tech sector which has transformed how financial services are delivered and consumed. This technological advancement offers significant opportunities for innovation and efficiency. Ethical Challenges: Despite the benefits, the introduction emphasizes that using big data and AI also raises significant ethical issues challenges [20].

MATERIALS AND METHOD

Data Sensitivity Score: A data sensitivity score is a measure of how sensitive or important a dataset is, based on the nature of the information it contains. It helps organizations classify data into levels such as low, medium, or high sensitivity. This score guides appropriate security and access controls, considering factors such as the presence of personal, financial, or health information, potential misuse, and compliance with regulations such as GDPR or HIPAA.

Privacy Risk Index: The Privacy Risk Index measures the likelihood and potential impact of privacy breaches related to the handling of personal data. It assesses factors such as the volume and sensitivity of the data collected, exposure to unauthorized access, and the effectiveness of privacy protections. Organizations use this index to identify vulnerabilities in data processes, ensure compliance with privacy laws, and ultimately reduce the risk of reputational damage or legal penalties resulting from data misuse.

AI Bias Metric: An The unbiasedness of machine learning models is assessed using the AI bias metric by measuring how their effects vary across population groups. It identifies differences such as unequal error rates or biased decision patterns that may result from skewed data or model design. These metrics help developers identify and mitigate bias in AI systems, and ensure that predictions or results are equitable and ethical across diverse populations and protected characteristics.

Ethical Compliance Score: The Ethical Compliance Score reflects how well an AI system or organization complies with ethical standards and policies. It considers factors such as transparency, accountability, data governance, user consent, and prevention of bias. This score helps businesses and

developers assess their adherence to ethical guidelines, legal obligations, and societal expectations. Higher scores indicate responsible technology use, while lower scores may indicate the need for improved ethical design and oversight in AI development and use.

Instructions for Machine Learning

Support Vector Regression: A potent machine learning method is support vector regression, especially for regression tasks. Its strengths include identifying support vectors, determining the optimal hyper plane, and efficiently handling high-dimensional data, making it a very versatile and useful tool. Known as "support vectors," the algorithm in SVR chooses important data points to define a hyper plane. The main goal is to maximize the margin between these vectors and minimize boundary violations. In essence, SVR seeks to place the hyper plane as near to the desired data points as feasible without going overboard. A multidimensional surface known as a hyper plane is utilized in SVR to represent the relationship between data points in feature space and to generate predictions. It looks as a straight line when there are two characteristics, but in more complicated situations, it transforms into a higher-dimensional surface. The basic notion is always the same, regardless of complexity: SVR seeks to identify the best hyper plane for precise predictions. It is possible to describe the compression problem as a typical quadratic programming task and solve it using regular quadratic programming techniques. These techniques are computationally intensive, though, particularly when the Gram matrix is too big to store in memory. A more effective substitute that cuts down on computing time and avoids memory overflow is the decomposition method.

Ada Boost Regression: Ensemble modelling was transformed by Freud and Schapiro's 1997 introduction of the Ada boost method. It has been widely applied to binary classification issues ever since. Ada boost transforms weak learners into strong and dependable models by combining many of them to increase prediction accuracy. Ada Boosting's basic concept is to train a model on a dataset initially, then iteratively create more models to account for the mistakes in the original ones. Until the prediction accuracy increases, this process keeps going. By combining several weak learners into a single, stronger model known as a "strong learner," Ada Boosting enhances performance. Ada Boost for predictive modelling is one of the machine learning algorithms that are accessible for distinct problem statements. Adaptive boosting, or Ada Boost, is a group method that enhances model performance. It is referred to as "adaptive" because it modifies the weights, giving misclassified occurrences greater weight in order to increase accuracy.

RESULT AND DISCUSSION

Four key metrics: Data Sensitivity Score (0-100): Measures how sensitive or confidential the data being processed is, with higher scores indicating more Sensitive data, including financial, health, or personal information. Privacy Risk Index (0-100): Assesses potential privacy breaches or data exposure risks, where higher values indicate greater threats to individual privacy. AI Bias Metric (0-1): Measures algorithmic bias and fairness issues, with values closer to 1 indicating more significant bias issues in decision-making or outputs. Ethical Compliance Score (0-100): Assesses adherence to ethical AI policies and regulatory standards, with higher scores indicating better compliance with ethical guidelines. The data shows considerable variation across all metrics, indicating diverse AI applications with varying risk profiles.

TABLE 1. Descriptive Statistics

	Descriptive Statistics			
	Data Sensitivity Score	Privacy Risk Index	Ai Bias Metric	Ethical Compliance Score
Count	500.0000	500.0000	500.0000	500.0000

Mean	49.4041	50.3368	0.4860	49.2595
Std	29.7697	15.2101	0.2866	16.7116
Min	0.0200	7.7200	0.0020	11.6700
25%	24.2525	39.2600	0.2300	36.2575
50%	47.8100	50.8800	0.4785	50.1700
75%	74.2800	61.6850	0.7313	60.4050
Max	99.9900	94.7300	0.9960	89.7700

Table 1 summarizes descriptive statistics for four key metrics across 500 observations. The average data sensitivity score and privacy risk index hover around 49 and 50, respectively, with wide variability, as shown by standard deviations of approximately 30 and 15. The AI bias metric averages near 0.49 on a scale up to 1, indicating moderate bias levels, while the ethical compliance score averages about 49, with values ranging broadly from 11.67 to nearly 90. Median and quartile values show data and risk scores are skewed, with higher values in the upper quartiles, suggesting significant variation in dataset sensitivity and compliance across samples.

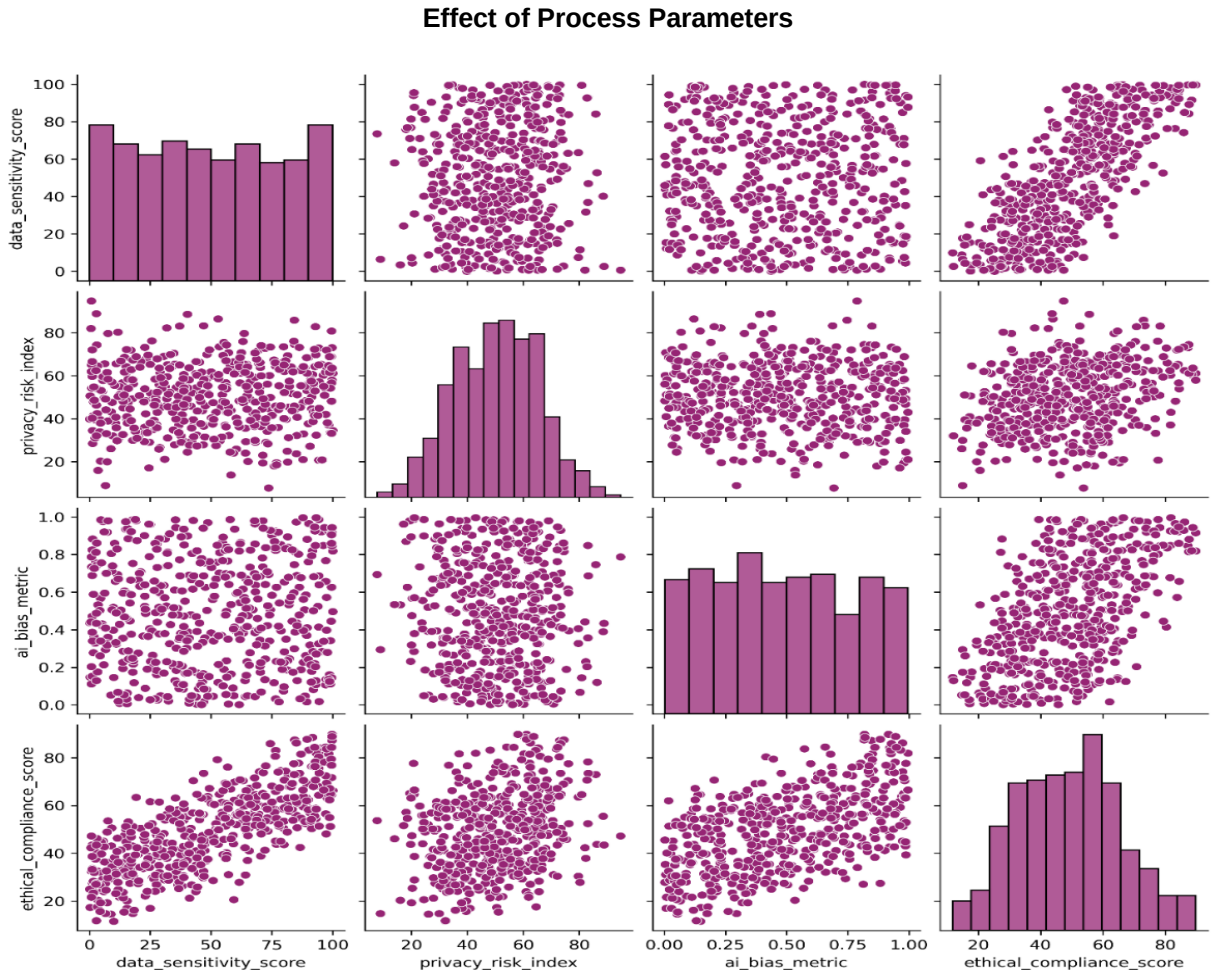

FIGURE 1. Scatter Plot of the Various Ethical Frameworks and Privacy Protection in AI-Driven Big Data

Figure 1 presents a scatter plot matrix exploring relationships between four variables: data sensitivity score, privacy risk index, AI bias metric, and ethical compliance score Regarding the use of big data driven by AI. The distribution of each variable is shown on the diagonal, and its pair wise relationships are shown in diagonal scatter plots relationships. Notably, ethical compliance score positively correlates with data sensitivity and inversely with AI bias. However, privacy risk appears relatively uncorrelated with the other variables. This suggests that improving ethical compliance may reduce AI bias and relate to handling more sensitive data, emphasizing the importance of integrated ethical and technical safeguards in AI systems.

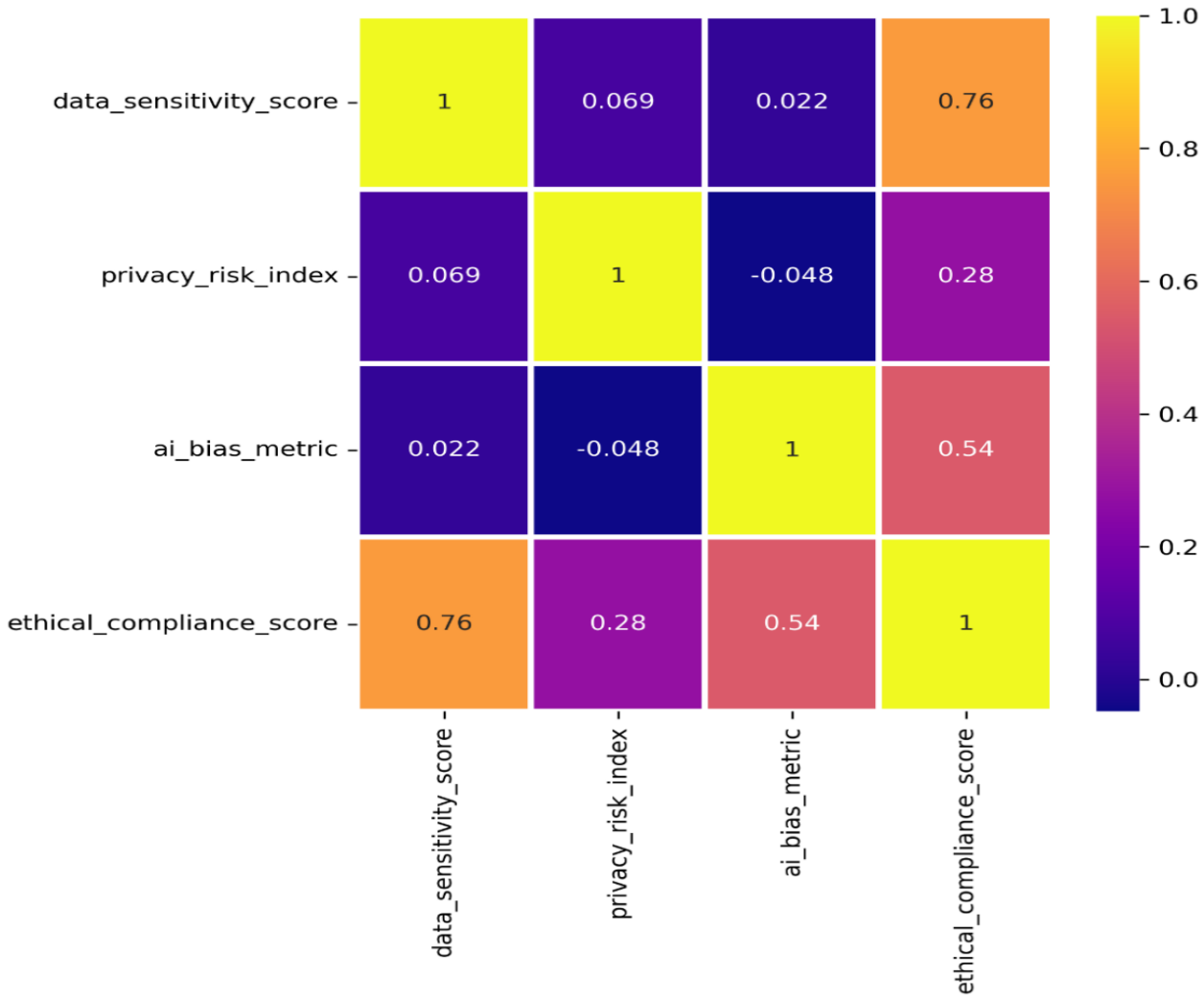

FIGURE 2. Correlation heat map on process parameters and effects

Figure 2 displays a correlation heat map illustrating the relationships among key parameters: data sensitivity score, privacy risk index, AI bias metric, and ethical compliance score. The most significant positive correlation (0.76) is observed between data sensitivity and ethical compliance, indicating that handling sensitive data tends to align with higher ethical standards. AI bias metric also correlates moderately (0.54) with ethical compliance, suggesting that addressing bias contributes to ethical AI.

Privacy risk index shows weak correlations with all variables, especially a slight negative correlation (-0.048) with AI bias.

Support Vector Regression

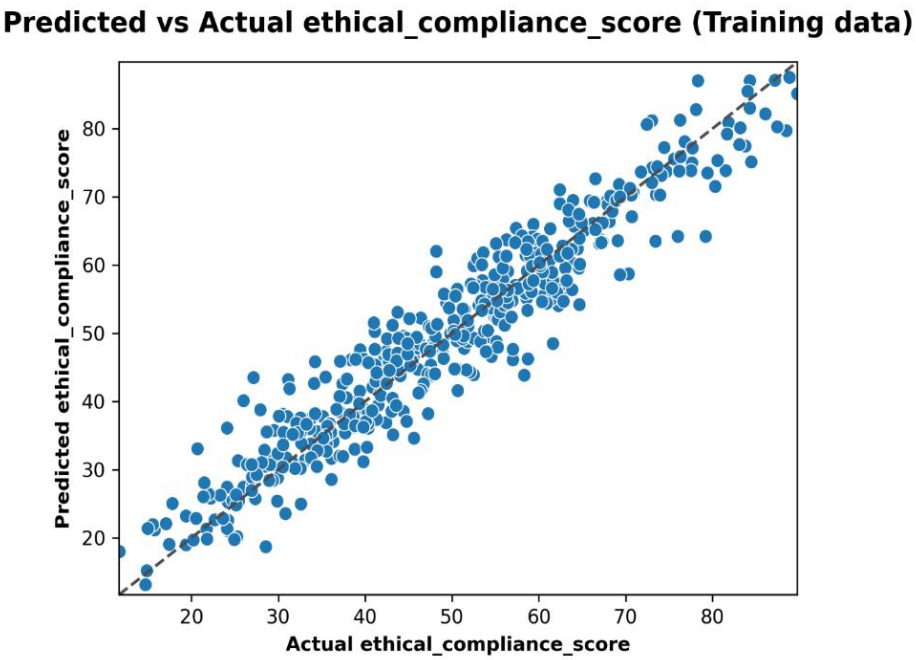

FIGURE 3. Support Vector Regression (Training data)

Figure 3 displays the predicted vs. actual ethical compliance score using Support Vector Regression on the training data. The dense clustering of points along the diagonal line suggests a strong fit and high prediction accuracy. The slight scatter indicates minor prediction errors, consistent with the high R^2 value (0.91) reported in Table 2.

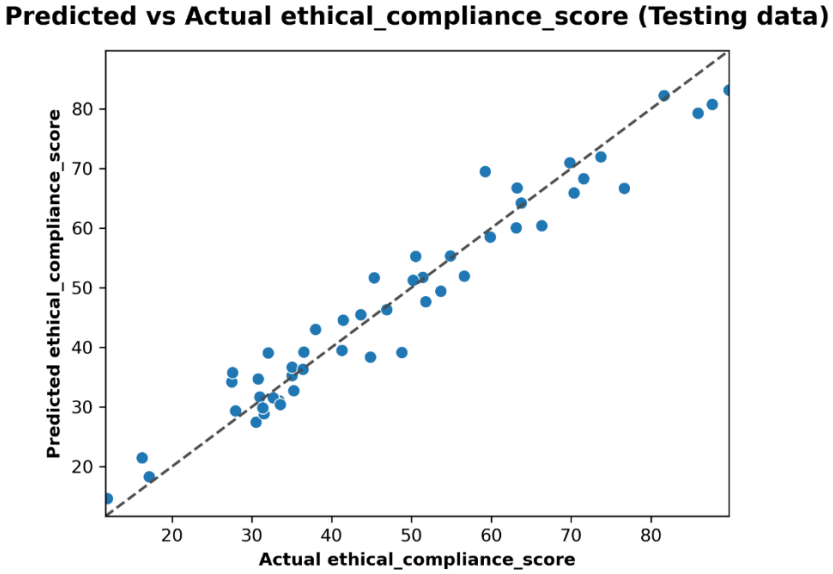

FIGURE 4. Support Vector Regression (Testing data)

Figure 4 shows the predicted vs. actual ethical compliance score for the testing data using Support Vector Regression. Data points closely follow the diagonal line, indicating high predictive accuracy. Minimal dispersion from the line reflects low error and excellent model generalization, consistent with the strong test performance metrics shown in Table 2.

Ada Boost Regression

FIGURE 5. Ada Boost Regression (Training data)

Figure 5 illustrates the predicted vs. actual ethical compliance score on training data using Ada Boost Regression. The points are tightly clustered along the diagonal, showing excellent model accuracy and strong correlation between predicted and actual values. Minimal scatter indicates low error rates and suggests the model fits the training data very well.

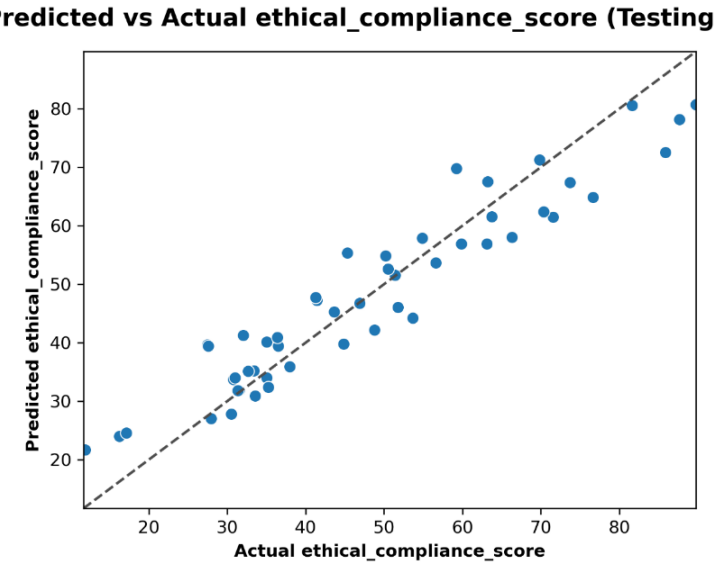

FIGURE 6. Ada Boost Regression (Testing data)

Figure 6 shows the predicted vs. actual ethical compliance score for the testing data using Ada Boost Regression. Most points align closely with the diagonal line, indicating strong prediction accuracy. Slight deviations suggest minor errors, but the model effectively captures the underlying pattern, reflecting a well-fitted and reliable predictive performance.

TABLE 2. Performance Metrics of Support Vector Regression (Training Data and Testing Data)

Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	Med AE
Train	SVR	0.91180	0.91188	23.71555	4.86986	3.75996	16.36542	0.01310	3.03107
Test	SVR	0.94526	0.94556	19.98395	4.47034	3.57005	10.31013	0.01088	3.06309

Table 2 indicates excellent Support Vector Regression performance. The high R^2 values (0.91 training, 0.95 testing) reflect strong model accuracy. Low MSE, RMSE, and MAE values confirm precise predictions. Testing metrics surpass training results, showing good generalization without overfitting. SVR delivers accurate and reliable regression outcomes across datasets.

TABLE 3. Performance Metrics of Ada Boost Regression (Training Data and Testing Data)

Data	Symbol	R2	EVS	MSE	RMSE	MAE	Max Error	MSLE	Med AE
Train	ABR	0.90343	0.90345	25.96667	5.09575	4.25663	11.73155	0.01564	3.87629
Test	ABR	0.88718	0.88719	41.18877	6.41785	5.26229	13.37054	0.02657	4.56311

Table 3 shows Ada Boost Regression's strong performance. The training R^2 of 0.90 and testing R^2 of 0.89 indicates high accuracy. Low MSE, RMSE, and MAE values across both sets suggest effective error minimization. Slightly higher errors in testing data reflect minor overfitting but overall robust predictive capability.

CONCLUSION

This comprehensive study provides significant insights into the complex relationship between ethics and privacy in AI-driven big data analytics, revealing important patterns with profound implications for the future of responsible AI development. The research demonstrates that ethical compliance in AI systems is not just a regulatory requirement, but a fundamental component that is strongly associated with data sensitivity handling and bias mitigation strategies. Statistical analysis of 500 observations across four key metrics reveals that organizations that process more sensitive data tend to maintain higher ethical standards, as evidenced by the strong positive correlation (0.76) between data sensitivity scores and ethical compliance. This finding suggests that the stakes in handling sensitive information naturally drive organizations toward more stringent ethical frameworks. However, the modest correlation (0.54) between AI bias metrics and ethical compliance indicates that while bias mitigation contributes to ethical AI, it represents only one component of a comprehensive ethical framework.

The machine learning models used in this study, especially support vector regression with its exceptional R^2 value of 0.95 on experimental data, demonstrate the potential to predict ethical compliance scores based on measurable parameters. This predictive ability opens up new avenues for proactive ethical monitoring and compliance assessment in AI systems, enabling organizations to identify potential ethical risks before they manifest as real-world harm. The relatively weak correlations observed with privacy risk indices across all variables highlight the complexity of privacy governance in AI systems. This suggests that privacy protection requires special attention and cannot be assumed to improve automatically through general ethical compliance measures. Therefore, organizations should develop dedicated privacy protections that operate independently of other ethical considerations. The implications of these findings

extend beyond academic interest in AI governance to practical applications. The research provides a foundation for developing automated ethical monitoring systems that can continuously assess AI applications against established ethical standards. For policymakers, these results underscore the importance of developing regulatory frameworks that address the multifaceted nature of AI ethics, encompassing not only bias and fairness, but also privacy protection and data sensitivity management. Future research should focus on expanding these metrics to include additional ethical dimensions such as transparency, accountability, and social impact. In addition, longitudinal studies that examine how ethical compliance evolves over time as AI systems mature will provide valuable insights into the dynamic nature of AI ethics. Integrating these findings into practical AI development workflows is an important step toward achieving the balance between technological innovation and the ethical responsibility demanded by society.

REFERENCES

1. Arora, Saransh, and Sunil Raj Thota. "Ethical considerations and privacy in AI-driven big data analytics." no. May (2024).
2. Shoghli, Arya, Mahsa Darvish, and Yasan Sadeghian. "Balancing Innovation and Privacy: Ethical Challenges in AI-Driven Healthcare." Journal of Reviews in Medical Sciences 4, no. 1 (2024): 1-11.
3. Khan, Wajahat Naseeb. "Ethical challenges of AI in education: Balancing innovation with data privacy." Journal of AI Integration in Education 1, no. 1 (2024): 1-13.
4. Bai, Ming, and Xiang Fang. "Ethical Considerations in Big Data-Enhanced Ai: A Comprehensive Analysis." EPH-International Journal of Educational Research 6, no. 3 (2022): 1-4.
5. Patel, Kaushikkumar. "Ethical reflections on data-centric AI: balancing benefits and risks." Available at SSRN 4993089 (2024).
6. Zhang, Yi, Mengjia Wu, George Yijun Tian, Guangquan Zhang, and Jie Lu. "Ethics and privacy of artificial intelligence: Understandings from bibliometrics." Knowledge-Based Systems 222 (2021): 106994.
7. Gustav, Lars Anders. "BALANCING PUBLIC SAFETY WITH PRIVACY CONCERNS ETHICAL IMPLICATIONS OF AI-POWERED."
8. Colson, Leonel, and Greyson Blake. "AI and Ethical Data Management: Protecting Privacy in Business Intelligence Systems." (2024).
9. Nabil, Ashrafur Rahman, Mahabub Sultan, Md Ruhul Amin, MD Nazim Akther, and Reaz Uddin Rayhan. "Ethical and Legal Considerations of AI in IT Project Management: Addressing AI Biases, Data Privacy, and Governance." Journal of Computer Science and Technology Studies 7, no. 2 (2025): 102-113.
10. Li, Yihong. "Security and Privacy of Artificial Intelligence with Ethical Concerns." In 2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC), pp. 660-667. IEEE, 2024.
11. Taneja, S., Shukla, R. P., & Singh, A. (2024). Navigating the Intersection of Ethics and Privacy in the AI Era (pp. 154–166). IGI Global. https://doi.org/10.4018/979-8-3693-2215-4.ch014.
12. Shukla, R. P., & Taneja, S. (2024). Ethical Considerations and Data Privacy in Artificial Intelligence (pp. 86–97). IGI Global. https://doi.org/10.4018/979-8-3693-2440-0.ch005
13. Rubel, Md., Emran, A. M., Borna, R. S., Saha, R. R., & Hasan, M. (2024). Ai-Driven Big Data Transformation and Personally Identifiable Information Security in Financial Data: A Systematic Review. 1(01), 114–128. https://doi.org/10.70008/jmldeds.v1i01.47
14. Pramanik, S. (2024). Managing the AI Period's Confluence of Security and Morality. Advances in Marketing, Customer Relationship Management, and e-Services Book Series, 131–146. https://doi.org/10.4018/979-8-3693-3478-2.ch006

15. Pandey, D., Jain, A., Suneetha, A., Gupta, P., Sriramakrishnan, G. V., Gopi, A. P., & Pramanik, S. (2024). Managing the AI Period's Confluence of Security and Morality (pp. 47–58). IGI Global. https://doi.org/10.4018/979-8-3693-2643-5.ch003
16. PK Kanumarlapudi, "AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning" Journal of Business Intelligence and Data Analytics., 2025, vol. 2, no. 2, pp. 1–17. doi: https://dx.doi.org/10.55124/jbid.v2i2.250
17. Aldboush, H. H. (2023). Building Trust in Fintech: An Analysis of Ethical and Privacy Considerations in the Intersection of Big Data, AI, and Customer Trust. International Journal of Financial Studies. https://doi.org/10.3390/ijfs11030090
18. Lacroix, P. (2019). Big Data Privacy and Ethical Challenges (pp. 101–111). Springer, Cham. https://doi.org/10.1007/978-3-030-06109-8_9
19. Ethical Considerations in AI-Enabled Big Data Research: Balancing Innovation and Privacy. (n.d.). International Journal of Control and Automation. https://doi.org/10.52783/ijca.v13i03.38350
20. Sanodia, G. (2021). AI Ethics and Big Data Governance in FinTech Applications. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-2269k
21. Unalp, Aynur. "Ethical Considerations in AI-Driven Cloud Solutions: Rethinking Business Intelligence for the Future." (2024).
22. Awad, Mariette, Rahul Khanna, Mariette Awad, and Rahul Khanna. "Support vector regression." Efficient learning machines: Theories, concepts, and applications for engineers and system designers (2015): 67-80.
23. Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14 (2004): 199-222.
24. Zhang, Fan, and Lauren J. O'Donnell. "Support vector regression." In Machine learning, pp. 123-140. Academic Press, 2020.
25. PK Kanumarlapudi. "Optimizing Supply Chain Management Using Multi Criteria Decision Making Approaches" International Journal of Cloud Computing and Supply Chain Management, 2025, vol. 1, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/ijccscm.v1i2.242
26. Solomatine, Dimitri P., and Durga L. Shrestha. "AdaBoost. RT: a boosting algorithm for regression problems." In 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541), vol. 2, pp. 1163-1168. IEEE, 2004.
27. Collins, Michael, Robert E. Schapire, and Yoram Singer. "Logistic regression, AdaBoost and Bregman distances." Machine Learning 48 (2002): 253-285.
28. PK Kanumarlapudi. "Improving Data Market Implementation Using Gray Relational Analysis in Decentralized Environments" Journal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 1, pp. 1–7. doi: https://dx.doi.org/10.55124/ jaim.v2i1.271

Integrating Big Data with Multi-Criteria Decision Making: COPRAS Method for Optimal Irrigation System Selection

Authors & Editors

Mr. Sudhakara Reddy Peram.

Correspondence

Engineering Leader, Illumio Inc., United States

Published By
Sciforce Publications
May 12, 2025

Integrating Big Data with Multi-Criteria Decision Making: COPRAS Method for Optimal Irrigation System Selection

Abstract: This study explores the integration in multi-criteria decision-making (MCDM) techniques with big data analytics has received significant attention in various fields. This combination enables more informed, Decision-making based on data insights using cutting-edge analytical techniques. A key MCDM method, complex proportionality estimation (COPRAS), is widely used in various domains, facilitating objective evaluations and optimal decision selection and its applications. As data volume and complexity increase, traditional analytical techniques struggle to effectively process and extract meaningful insights. This research explores how advanced computational approaches can mitigate these Big data analytics presents both challenges and opportunities in improving decision-making processes. While dealing with vast and complex datasets is challenging, advanced analytical techniques help extract valuable insights, improving the accuracy and efficiency of decision-making.

By leveraging these capabilities, organizations can make more informed and strategic choices across a variety of domains has advanced from dealing with structured datasets to managing vast amounts of unstructured information, particularly in fields that require rapid responses, such as agricultural pest management, climate modeling, and biodiversity assessment. The review highlights the significant data volumes generated Remote sensing applications face difficulties due to the huge volume of image data; however, technologies such as fog computing help solve these problems by enabling decentralized processing and reducing latency. IoT and cloud computing have greatly improved data management and processing capabilities management and processing capabilities. Processing capabilities across industries. The COPRAS method stands out as an effective multi-criteria decision-making tool that helps evaluate alternatives by combining positive and negative ideal solutions. Its computational power and structured algorithm make COPRAS particularly useful for solving complex selection problems involving conflicting criteria.

This study highlights the versatility of this method through its successful application in various fields, including research assistant performance evaluation, supplier selection, construction method evaluation, educational institution ranking, and irrigation system selection. In addition, the research explores how COPRAS can be integrated with other analytical frameworks, Techniques such as the Fuzzy Analytical Hierarchy Process (AHP) and TOPSIS improve decision making by providing a structured approach to evaluating alternatives by providing structured and systematic evaluation methods systematic and structured approaches to evaluating multiple criteria qualities of decision-making by reducing uncertainty and accommodating subjective judgments. These hybrid approaches leverage complementary analytical strengths, providing more robust solutions to complex decision-making situations. As organizations increasingly adopt data-driven strategies in dynamic environments, this study contributes to understanding how advanced analytical techniques can Transform raw data into meaningful insights and enable more informed and strategic decision-making improving outcomes, COPRAS and its hybrid applications are proving valuable in both technical and managerial domains.

Keywords: COPRAS Method, Big data analytics, multi-criteria decision making (MCDM), irrigation systems, fuzzy Logic.

INTRODUCTION

Exceptions include projects related to animals and weeds, as immediate action is required to detect weeds and diseases. In addition, exceptions include efforts focused on Weather and climate change analysis, biodiversity assessment and agricultural decision-making rely on extensive datasets from multiple sources to generate predictions, model environmental changes, and support farmers in their operations. Among these, remote sensing applications handle the largest data volumes due to the significant amount of image-based information. [1] Most studies deal with relatively moderate to low speeds and medium to high data volumes with diversity. However, there are exceptions include projects related to animal and weed management, which require high-speed data processing due to the urgency of identifying weeds and diseases. Similarly, Weather and climate change research, biodiversity assessments, and agricultural decision-making applications show high data heterogeneity because they rely on multiple sources for weather forecasting, climate change modelling, and biodiversity assessment, and support agricultural operations. [2] This study focuses only on Twitter as a source of big data and does not aim to address all the challenges commonly encountered in big data analysis.

However, unlike Hadoop, which processes data in batches, it will focus on handling a series of tweets. This review paper will explore big data analysis and related tools, particularly Apache Storm, and will include a comparison of available tools with Apache Storm and its justification. One section will cover materials and methods, another section will present the results of big data analysis using Storm, and the final section will discuss the findings and outline future work. [3] These characteristics make extracting valuable insights from big data a very challenging task. A simple analogy is a group of blind people This is like Trying to describe a large elephant (see Figure 1), this represents the concept of big data. Each blind person forms an opinion about the elephant based only on the specific part they can touch. Since their perspectives are limited to limited areas, it is not surprising that their conclusions differ depending on the part they encounter - one person might describe the elephant as a rope, another as a pipe, and another as a wall. [4]

This research examines how network relationships between different news sources shape the agendas of different stakeholders, including the media, the public, and interest groups. It seeks to advance The NAS model is extended by a systematic Exploring the impact of fake news on media agenda-setting across different media platforms. This study specifically examines how media organizations create interconnected issue agendas, linking different topics to influence public perception, and how these issue networks evolve across different media channels landscapes. [5] This paper presents A hierarchical, decentralized fog computing architecture is proposed for big data analytics in smart cities. Considering the geo-distributed nature of data generated by a large number of sensors, a multi-layer fog computing architecture is deployed at the intelligent edge of the network. Computing nodes in each layer process latency-sensitive applications and implement fast control loops to improve the security of critical infrastructure.

The effectiveness of this fog-based computing paradigm is demonstrated through a smart pipeline monitoring case study is developed that demonstrates its performance and feasibility for future city-scale implementation. [6] In recent years, advances in the field of Internet of Things (IoT) and big data, advanced technologies support the collection, processing, and analysis of vast datasets, enabling real-time decision-making and improving overall system performance analytics has attracted significant attention and has become essential technologies, driving the advancement of applications in many domains. Sensor technology has also seen significant progress, leading to the development of various specialized sensors, such as environmental and gas sensors, tailored to specific applications. [7] Expanding Building on the NAS model and the concept of issue ownership, this article introduces a new framework called the issue ownership network. This approach highlights the influential role of the media and political campaigns in shaping public opinion and debates perceptions by shaping the interactions between political parties or candidates and interconnected issue network's opinion by linking political parties or candidates not only to individual issues but also to interconnected issue networks. Within these

networks, certain issues act as bridges, connecting a party or candidate to additional topics. This This theoretical approach broadens the understanding This perspective broadens the understanding of both the NAS model and publishing rights theory. [8] According to world systems theory (WST), countries are classified into three interconnected zones based on Occupying an intermediate position between the two, they balance elements of both high-skill and low-skill production. The contrast between the Peripheral regions is often exploited because capitalist countries dominate underdeveloped regions regions act as an intermediary between these two groups. The imbalance between the imbalance between the dominant capitalist countries and the underdeveloped regions often leads to the exploitation of the peripheral regions and their resources semi-peripheral regions, further reinforcing the hegemonic dominance of the core. [9]

This article explores Big data introduces both challenges and opportunities The shift from mass communication to networked communication, to systematic and quantitative content analysis in media and communication research, has led to a vast increase in publicly available online content, surpassing the scope of traditional content analysis methods to networked communication has dramatically expanded the amount of publicly available online content, creating massive datasets beyond the limits of traditional analysis.

As a result, media content analysts are adopting innovative approaches to address a series of challenges in text selection and coding. [10] Since the emergence of computer science, the concept of "Big Data" has been a fundamental aspect of the field significant topic. Initially, it referred to the volume of data that exceeded the processing capabilities of traditional database methods and tools. As new storage technologies emerged, data access increased dramatically, leading to an exponential growth in the amount of information available. While the initial focus was Researchers and practitioners who initially focused on structured data later realized that most of the world's information exists in extensive, unstructured datasets forms, consisting primarily of text and images. [11]

Thanks to GEE’s capabilities in overcoming the challenges of big data analytics, especially its ability to quickly process large datasets within short timeframes, plays a key role in remote sensing for geographic analysis at both local and global scales. As previously mentioned, remote sensing big data analytics incorporates both intrinsic and extrinsic properties. A review of studies using Google Earth Engine (GEE) as a geospatial big data analytics tool highlights research efforts in this area exploring these applications uses multi-temporal data to improve analytical capabilities track events significantly outnumbers those using single-date data, highlighting the importance of the intrinsic features of geospatial big data. [12] For small data, analysis was performed by randomly selecting sample subsets considered representative of the overall dataset entire dataset.

However, since only partial data was analyzed, the extracted information was often inaccurate and incomplete, leading to suboptimal results and poor performance. This is especially true in Real-time network analysis and troubleshooting rely on accurate and timely information to ensure effective problem resolution and optimal performance accurate solutions. Such results can only be achieved through the analysis of the entire dataset or big data. [13] Despite the growing enthusiasm the growing interest in "big data" stems from its considerable operational and strategic potential understanding of what the concept actually entails are still limited. As a result, many potential adopters of ‘big data’ face challenges in fully understanding the concept and unlocking its business value. Limited empirical studies have assessed the true potential of “big data”. This study aims to address this knowledge gap building on existing big data research and providing a detailed case study of an Australian state emergency service. [14]

New technologies and methods need to be developed to identify factors associated with crashes, identify dangerous times and places, and thereby aid traffic crash Using data analytics in decision-making processes, this paper introduces a new big data analytics platform for analyzing road accidents in the UK. By integrating Using data mining and deep learning techniques, this study initially presents the data effectively through an interactive map enhances analysis and interpretation highlights crash hotspots based on time and location. [15]

MATERIALS AND METHOD

Brand: A brand is a product, service, or concept that is uniquely identified from others in the marketplace, allowing it to be easily communicated and marketed in general. Branding refers to the process of creating and enhancing a brand name, its characteristics, and its personality.

Price: At its core, price is the amount of money required to purchase a good or service. However, its role extends beyond just monetary value; it also reflects the perceived value of a good by consumers and indicates how much they are willing to pay for it.

Quality: In general, "quality" refers to the degree or degree of excellence of something, usually indicating a high standard. It can be used to describe the character of a product, service, or a person.

Product Type: Product types refer to the different types of goods that businesses produce and sell to consumers or other organizations. For example, businesses may create consumer goods for individuals, while industrial goods are often targeted at other organizations and contribute to supporting the production process.

Durability: Technically, durability refers to the length of time a product can be used before it begins to deteriorate. Alternatively, it can be defined as the length of time a product can be used before it becomes so damaged that replacement is more practical than constant repair.

Sprinkler Irrigation: Sprinkler irrigation is best for various field conditions, except for dense clay soils. In this method, water is distributed evenly and efficiently. As a result, water is used carefully and effectively.

Drip Irrigation: Sprinkler systems have pipes buried in the ground that spray water onto plants with heads above the surface, while drip irrigation uses pipes that run along the ground, gradually releasing water directly into the soil around the plants.

Pivot Irrigation: Sprinkler systems involve pipes installed underground, with heads above ground that spray water onto plants, whereas drip irrigation uses pipes located close to the ground, slowly releasing water directly into the soil around the plants.

Linear Irrigation: Linear motion irrigation It is an automatic irrigation system that delivers water in a direct path to the crops, making it best suited for rectangular or square fields. The system consists of multiple hose extensions, usually mounted on wheeled towers, that travel back and forth across the field to ensure even water distribution.

COPRAS method: In this study, the performance of a research assistant is evaluated using the COPRAS method based on various criteria, including undergraduate GPA, master's degree GPA, PhD GPA, foreign language score, time to complete master's and PhD courses, number of conferences attended, and number of articles written. As a result, the research assistant labeled x1 is determined to have the highest performance score. Upon further examination, this individual stands out with a higher undergraduate GPA, a higher number of conferences attended, and a higher foreign language score; however, they have written the lowest number of articles. [16] The locations of workplaces were determined using data generously provided by the Social Insurance Information Service. Since this data is very dynamic and subject to frequent changes, its accuracy is low compared to data on places of residence. In addition, some assumptions and indirect analysis methods were used in the calculations. Information on the main public attraction centers, along with visitor data for these centers, was obtained from a newly developed special plan outlining the distribution of large markets in the city of Kaunas. Another important dataset identified by experts as significant in a small city concerns the level of development of the public transport network. [17] This paper presents a comprehensive fuzzy multi-criteria decision-making model that combines the fuzzy analytic hierarchy process (AHP) with the complex proportionality rating (COPRAS) method to evaluate the relative performance of IITs. The main

contribution of this approach is to improve the evaluation accuracy by leveraging the strengths of both techniques. The model is distinguished by its robustness, ease of implementation, and minimal reliance on complex mathematical calculations. In addition, the evaluation criteria are aligned with the preferences of stakeholders to ensure relevance. The fuzzy AHP method is used to determine the relative importance of these criteria. [18] The COPRAS method evaluates alternatives by considering both positive and negative ideal solutions.

The optimal choice is determined by having the highest positive ideal solution value and the lowest negative ideal solution value. In addition, this method improves decision making by providing a structured evaluation framework that includes normalization of alternative data to eliminate computational inconsistencies. In this context, this study integrates the fuzzy AHP-COPRAS method as a decision support system for selecting new student admissions in Matson Bang Kalan. [19] Stable and uninterrupted cash flow is A key factor in the smooth execution of management activities. While companies can implement internal strategies to directly impact Production quality, sales potential and even raw material costs small disruptions in the collection process can have significant impacts process - often dependent on external factors - can derail future plans and disrupt cash flow. Maintaining stable cash flow is essential for businesses to develop prudent and future-oriented strategies. [20] Although many studies have investigated supplier selection using various MCDM methods, a structured and systematic mathematical approach is still needed to address the uncertainties in the selection process. This paper advances the COPRAS method, a modern MCDM technique designed to handle decision-making challenges involving conflicting and incomparable criteria.

The proposed approach adapts COPRAS to support fuzzy multi-criteria group decision-making within the framework of interval-type fuzzy sets. [21] The COPRAS method, first proposed by Saatsakis et al., is used first for rank alternatives based on multiple criteria and their relative weights. It follows a step-by-step approach to evaluate and prioritize alternatives according to their importance and application levels. Its simplicity has made it a widely used method in various fields, including construction, material selection, and contractor evaluation. [22] To establish a reliable framework, this study combines the outputs of the fuzzy TOPSIS and COPRAS methods. Combining multiple solution approaches improves the quality of decision-making in multi-criteria analysis and provides a strong foundation for accurate results. This framework It includes various scaling techniques are used, such as COPRAS, ANP, AHP, and TOPSIS are widely used scaling techniques. In particular, the COPRAS method evaluates alternatives by analyzing their positive and negative outputs. Notably, as indicated in the portfolio section, banking is excluded from consideration. [23] The main principle of both methods is that the optimal The Optimal an alternative should be positioned as close as possible to the positive ideal solution, while maintaining a significant distance from the negative ideal solution. However, the ranking results of the HFL-COPRAS method differ from those of the HFL-TODIM method. This difference stems primarily from the fact that HFL-TODIM is based on expectation theory, which is fundamentally different from the approach used in HFL-COPRAS.

Furthermore, discrepancies are observed in the ranking results of HFL-COPRAS and those of other methods, including HFL-TOPSIS. [24] In addition, Yu introduced a new concept of determining the decision maker's weights the mean is determined by Both negative and positive ideal solutions are taken into account the evaluation process. In addition, COPRAS method has gained significant popularity in various engineering and management fields in recent years because it effectively evaluates multiple criteria while considering both maximization and minimization objectives. In view of this The COPRAS method can be improved by effectively integrating the mean, negative, and positive ideal solutions. [25] To maintain a continuous production flow, it is necessary to obtain the necessary components and materials on time while optimizing the costs associated with this logistics subsystem. The focus of this research is a company that produces pre-insulated pipes and requires steel procurement. To evaluate suppliers, this paper uses a combination of multi-criteria analysis techniques, using the Analytic Hierarchy Process (AHP) to determine the relative importance of criteria. [26] The aim of this research is to identify the most effective construction method using cold-formed steel structures to rebuild areas affected by

natural disasters. The process begins by determining key criteria through interviews and a conference meeting. Next, the AHP method is used to assess the importance of each criterion. Finally, the COPRAS-G method is used to evaluate the top three approaches to implementing cold-formed steel structures. The main research steps are outlined as follows. [27] In recent years, the rapid increase in construction projects has made selecting the right project managers more important. As different projects require unique skills and abilities, identifying suitable project managers has become a major challenge in project implementation. Partners, consultants, and contractors are looking for highly qualified project managers who are in short supply. Despite the fact that skilled project managers can often negotiate their own salaries, even specialized recruitment agencies struggle to find suitable candidates. [28] The “Materials and Methods” section describes the methodology of the study and provides theoretical insights into the Structural Equation Modeling (SEM) and COPRAS methods used in the research.

The “Application of Structural Equation Modeling (SEM)” section explains the implementation process of SEM. The “Results and Discussion” section presents the findings and conclusions derived from the SEM and COPRAS data, including the selection of renewable energy sources using the COPRAS method. Finally, the Conclusion section summarizes the key insights and final conclusions. [29] The study used the entropy method to determine the importance of criteria in the decision-making process. The results ranked energy cost, labor requirements, and price as the most important factors in that order. To evaluate irrigation system alternatives, the research used the COPRAS method, which is a multi-criterion decision-making (MCDM) approach. The findings indicated that linear irrigation systems were the most suitable choice for the study area. Rising input costs have prompted agricultural companies to adopt cost-saving strategies.

RESULT AND DISCUSSION

TABLE. 1 Big data analysis

	Sprinkler Irrigation	**Drip Irrigation**	**Pivot Irrigation**	**Linear Irrigation**
Brand	0.19000	0.55000	0.11000	0.63000
Price	0.16000	0.27000	0.23000	0.25000
Quality	0.12000	0.34000	0.99000	0.15000
Product Type	0.15000	0.97000	0.32000	0.45000
Durability	0.18000	0.44000	0.76000	0.22000

Table 1 shows Findings from a big data analysis conducted using COPRAS method to evaluate various irrigation systems: sprinkler irrigation, drip irrigation, centerline irrigation, and linear irrigation. The evaluation is based on five criteria: brand, price, quality, product type, and durability. Linear irrigation scores highest in the brand category, while drip irrigation excels in product type and price. Centerline irrigation stands out in quality and durability, highlighting its strong performance in these areas. Sprinkler irrigation scores low on most criteria, suggesting that it may not be as optimal as other systems in these factors.

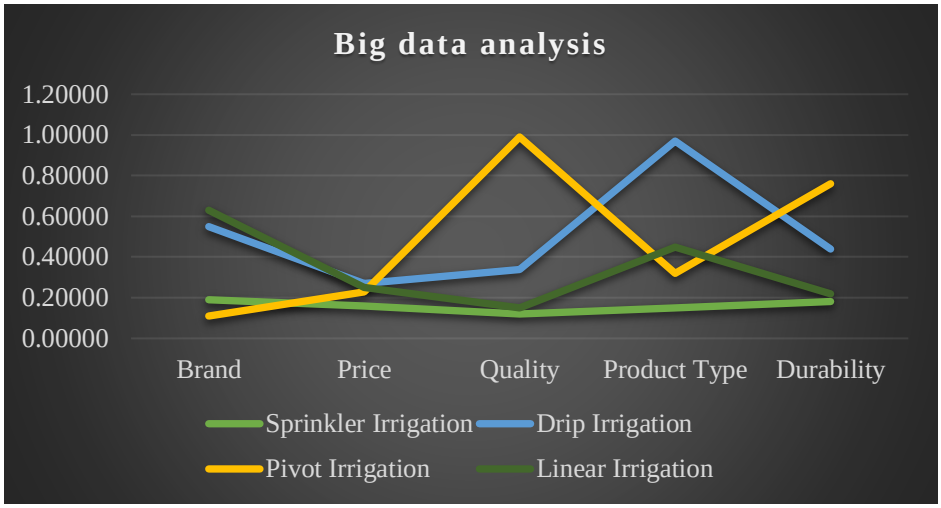

FIGURE. 1 Big data analysis

Figure 1 Using and displaying the results of big data analysis COPRAS method that compares four irrigation systems: sprinkler, drip, pivot, and linear irrigation. The analysis evaluates the systems based on five criteria: brand, price, quality, product type, and durability. Linear irrigation excels in brand, drip in product type, and pivot in quality and durability.

TABLE. 2 Normalized Data

	Normalized Data			
	Sprinkler Irrigation	**Drip Irrigation**	**Pivot Irrigation**	**Linear Irrigation**
Brand	0.2375	0.2140	0.0456	0.3706
Price	0.2000	0.1051	0.0954	0.1471
Quality	0.1500	0.1323	0.4108	0.0882
Product Type	0.1875	0.3774	0.1328	0.2647
Durability	0.2250	0.1712	0.3154	0.1294

Table 2 shows normalized data from the COPRAS method used to evaluate four irrigation systems: sprinkler, drip, pivot, and linear irrigation. The normalized scores indicate the performance of each system on five criteria: brand, price, quality, product type, and durability. Linear irrigation performs best on brand, while drip irrigation excels on product type. Pivot irrigation stands out on quality and durability. Sprinkler irrigation generally receives low normalized scores on most criteria, indicating its weak performance compared to other systems. This normalized data helps identify the most reliable and optimal irrigation system for specific needs.

TABLE. 3 Weight

	Weight			
Brand	0.95	0.95	0.95	0.95
Price	0.95	0.95	0.95	0.95
Quality	0.95	0.95	0.95	0.95
Product Type	0.95	0.95	0.95	0.95
Durability	0.95	0.95	0.95	0.95

Table 3 shows The COPRAS method assigns weights to each criterion for evaluating four irrigation systems: sprinkler, drip, pivot, and linear irrigation. The five criteria – brand, price, quality,

product type, and durability – each have an equal weight of 0.95 across all systems. This indicates that all criteria are considered equally important in the evaluation process all criteria are considered equally important and influential in the evaluation process. The equal distribution of weights ensures that no particular factor is given more priority than the others, resulting in a fair and balanced evaluation of irrigation systems based on the five criteria.

TABLE. 4 Weighted normalized decision matrix

	Weighted normalized decision matrix			
Brand	0.23	0.20	0.04	0.35
Price	0.19	0.10	0.09	0.14
Quality	0.14	0.13	0.39	0.08
Product Type	0.18	0.36	0.13	0.25
Durability	0.21	0.16	0.30	0.12

Table 4 shows the weighted and normalized result matrix obtained from the COPRAS method used to evaluate four irrigation systems: sprinkler, drip, pivot, and linear irrigation. This matrix shows the weighted scores of each system on five criteria: brand, price, quality, product type, and durability. Linear irrigation performs best on brand, while drip irrigation leads on product type. Pivot irrigation excels on quality and durability. Sprinkler irrigation generally scores low on most criteria, indicating that it is less favorable compared to other systems. The matrix provides a complete comparison of irrigation systems based on their respective performance.

TABLE. 5 Bi, Ci

	Bi	**Ci**
Brand	0.429	0.395
Price	0.290	0.230
Quality	0.268	0.474
Product Type	0.537	0.378
Durability	0.376	0.423

Table 5 shows the \ (B_i \) and \ (C_i \) values obtained from the COPRAS method, which are used to evaluate four irrigation systems based on five criteria: brand, price, quality, product type, and durability. In this context, \ (B_i \) denotes The positive ideal solution is denoted by \ (B_i \), while \ (C_i \) denotes the negative ideal solution. for the Brand criterion, \ (B_i \) value 0.429 and Ci is 0.395, indicating a more balanced evaluation. In contrast, for quality, the Ci value (0.474) is higher than the Bi value (0.268), indicating a closer approximation to the negative ideal solution. These values allow us to compare the performance of each criterion with ideal and non-ideal solutions.

TABLE. 6 Min(Ci)/Ci

	Min(Ci)/Ci
Brand	0.5826
Price	1.0000
Quality	0.4859
Product type	0.6101
Durability	0.5452

Table 6 presents the minimum(\(C_i\))/\(C_i\) values obtained from the COPRAS method, which indicate the relative distance of each criterion's performance from the negative ideal solution. These values estimate how closely each irrigation system approaches the worst-case scenario. For price, a value of 1.0000 indicates that it is very far from the negative ideal solution, indicating a very favorable outcome. On the other hand, quality has a value of 0.4859, indicating that it is close to the negative ideal

solution. These values help estimate how much each criterion deviates from the worst-case performance, providing insight into the relative performance of the systems.

TABLE. 7 Qi, Ui

	Qi	Ui
Brand	0.772	86.1704
Price	0.879	98.0953
Quality	0.555	61.8780
Product type	0.896	100.0000
Durability	0.698	77.8508

When analyzing the given data using the COPRAS method, we see that the “product type” criterion has the highest Qi value (0.896) and is considered the most significant (Ui = 100.0000). “Price” comes next, with Qi = 0.879 and Ui = 98.0953, highlighting its important role in decision-making. “Brand” is in third place, with Qi = 0.772 and Ui = 86.1704. “Durability” (Qi = 0.698, Ui = 77.8508) has a moderate impact. Meanwhile, “Quality” ranks the lowest (Qi= 0.555, Ui = 61.8780), making it the least influential factor. This analysis indicates that product type and price significantly shape evaluations, while quality has a minor effect.

TABLE: 8 Rank

	Rank
Brand	**3**
Price	**2**
Quality	**5**
Product type	**1**
Durability	**4**

By analyzing the ranking data with the COPRAS method, we see that “product type” holds the first place (rank 1), making it the most important factor in the rating. “Price” is in second place, emphasizing its strong impact on decision-making. “Brand” receives the third rank, indicating moderate importance. “Durability” is in fourth place, showing that while it contributes to the rating, it has less influence than the higher ranked criteria. “Quality” is in last place (5), indicating that it has the least effect on the rating. This ranking analysis highlights that product type and price are the most influential factors, while quality plays the least role.

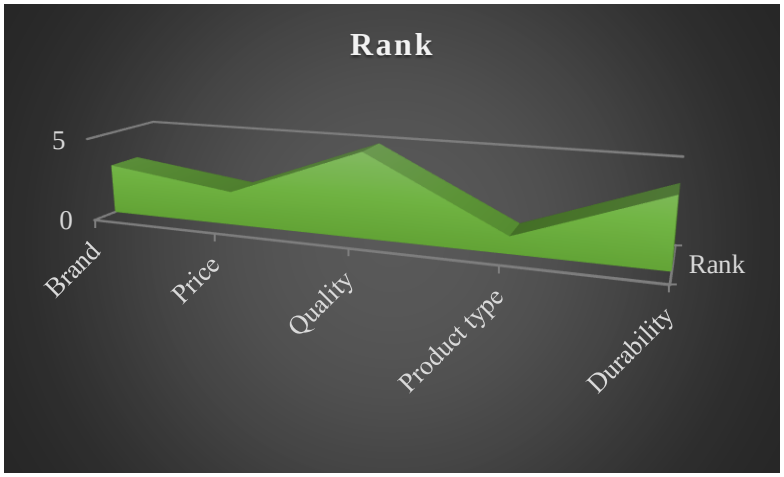

FIGURE. 2 Rank

Analyzing the rankings from Figure 2 using the COPRAS method, “product type” takes the top spot, making it the most important factor. “Price” is in second place, highlighting its importance. “Brand” takes third place, and “Durability” takes fourth place. “Quality” is in last place (fifth), indicating its minimal impact on the evaluation process.

CONCLUSION

A thorough review of the literature provides key insights into the extensive use of big data analytics multi-criteria decision-making methods are widely used in various fields, including COPRAS method being a widely recognized approach. Big data analysis introduces both challenges and opportunities, reshaping industries by shifting from traditional data processing to managing massive, unstructured datasets. Remote sensing applications generate enormous data volumes due to image size, while exceptions include animals, weeds, weather forecasting, climate modeling, biodiversity assessment, and agricultural decision-making, which require real-time processing. The hierarchical decentralized fog computing architecture offers the potential for smart city applications by efficiently handling geo-distributed sensor data. Similarly, the integration of IoT technologies with big data analytics has significantly improved various sectors, including agriculture through sensor-based improvements. The COPRAS (complex proportional estimation) method It has proven its effectiveness as a valuable tool balancing multi-criteria analysis, positive and negative best solutions with computational efficiency.

The literature highlights its applications in evaluating research assistant performance, supplier selection, disaster recovery construction methods, renewable energy source selection, irrigation system evaluation, and educational institution performance. Combining this method with other analytical approaches, for example the fuzzy analytic hierarchy process (AHP), enhances its ability to manage uncertainty and subjective judgments. This integrated framework improves decision quality by providing a systematic mathematical approach that effectively addresses ambiguity in complex choice problems. In agricultural applications, COPRAS is used to evaluate irrigation systems, identifying linear irrigation as the most suitable option for rectangular fields based on factors such as energy costs, labor requirements, and price - key considerations as rising input costs push agricultural organizations toward efficiency. When organizations navigate complex decisions involving multiple, often conflicting criteria, the COPRAS method provides a structured and transparent approach that balances both maximization and minimization objectives. Its successful implementation in a variety of sectors highlights its adaptability and effectiveness. Ultimately, as data volumes grow and decision-making environments become increasingly complex, enhancing big data analytics with robust multi-criteria methods like COPRAS will be critical to gaining insights, evaluating alternatives, and facilitating informed Decision-making in many industries.

REFERENCES

1. Kamilaris, Andreas, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldú. "A review on the practice of big data analysis in agriculture." Computers and electronics in agriculture 143 (2017): 23-37.
2. Shoro, Abdul Ghaffar, and Tariq Rahim Soomro. "Big data analysis: Ap spark perspective." Global Journal of Computer Science and Technology: C Software & Data Engineering 15, no. 1 (2015): 7-14.
3. Iqbal, Muhammad Hussain, and Tariq Rahim Soomro. "Big data analysis: Apache storm perspective." International journal of computer trends and technology 19, no. 1 (2015): 9-14.
4. Wu, Xindong, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. "Data mining with big data." IEEE transactions on knowledge and data engineering 26, no. 1 (2013): 97-107.
5. Vargo, Chris J., Lei Guo, and Michelle A. Amazeen. "The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016." New media & society 20, no. 5 (2018): 2028-2049.

6. Tang, Bo, Zhen Chen, Gerald Hefferman, Tao Wei, Haibo He, and Qing Yang. "A hierarchical distributed fog computing architecture for big data analysis in smart cities." In Proceedings of the ASE BigData & SocialInformatics 2015, pp. 1-6. 2015.
7. Channe, Hemlata, Sukhesh Kothari, and Dipali Kadam. "Multidisciplinary model for smart agriculture using internet-of-things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis." Int. J. Computer Technology & Applications 6, no. 3 (2015): 374-382.
8. Guo, Lei, and Chris Vargo. "The power of message networks: A big-data analysis of the network agenda setting model and issue ownership." Mass communication and society 18, no. 5 (2015): 557-576.
9. Guo, Lei, and Chris J. Vargo. "Global intermedia agenda setting: A big data analysis of international news flow." Journal of Communication 67, no. 4 (2017): 499-520.
10. Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. "Content analysis in an era of big data: A hybrid approach to computational and manual methods." Journal of broadcasting & electronic media 57, no. 1 (2013): 34-52.
11. Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. "Big data: Issues and challenges moving forward." In 2013 46th Hawaii international conference on system sciences, pp. 995-1004. IEEE, 2013.
12. Tamiminia, Haifa, Bahram Salehi, Masoud Mahdianpari, Lindi Quackenbush, Sarina Adeli, and Brian Brisco. "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review." ISPRS journal of photogrammetry and remote sensing 164 (2020): 152-170.
13. Parwez, Md Salik, Danda B. Rawat, and Moses Garuba. "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network." IEEE Transactions on Industrial Informatics 13, no. 4 (2017): 2058-2065.
14. Wamba, Samuel Fosso, Shahriar Akter, Andrew Edwards, Geoffrey Chopin, and Denis Gnanzou. "How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study." International journal of production economics 165 (2015): 234-246.
15. Feng, Mingchen, Jiangbin Zheng, Jinchang Ren, and Yanqin Liu. "Towards big data analytics and mining for UK traffic accident analysis, visualization & prediction." In Proceedings of the 2020 12th International conference on machine learning and computing, pp. 225-229. 2020.
16. Organ, Arzu, and Engin Yalçın. "Performance evaluation of research assistants by COPRAS method." European Scientific Journal 12, no. 10 (2016): 102-109.
17. Zagorskas, Jurgis, Marija Burinskienė, Edmundas Zavadskas, and Zenonas Turskis. "Urbanistic assessment of city compactness on the basis of GIS applying the COPRAS method." Ekologija 53, no. 2 (2007): 55-63.
18. Das, Manik Chandra, Bijan Sarkar, and Siddhartha Ray. "A framework to measure relative performance of Indian technical institutions using integrated fuzzy AHP and COPRAS methodology." Socio-Economic Planning Sciences 46, no. 3 (2012): 230-241.
19. Kustiyahningsih, Yeni, and Ismy Qorry Aini. "Integration of FAHP and COPRAS method for new student admission decision making." In 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), pp. 1-6. IEEE, 2020.
20. Özbek, Aşir, and Emel Erol. "Ranking of factoring companies in accordance with ARAS and COPRAS methods." International Journal of Academic Research in Accounting, Finance and Management Sciences 7, no. 2 (2017): 105-116.
21. Keshavarz Ghorabaee, Mehdi, Maghsoud Amiri, Jamshid Salehi Sadaghiani, and Golnoosh Hassani Goodarzi. "Multiple criteria group decision-making for supplier selection based on COPRAS method with interval type-2 fuzzy sets." The International Journal of Advanced Manufacturing Technology 75 (2014): 1115-1130.

22. Esbouei, Saber Khalili, and Abdolhamid Safaei Ghadikolaei. "Applying FAHP and COPRAS methods for evaluating financial performance." International Journal of Managment, IT and Engineering 3, no. 11 (2013): 10-22.
23. PK Kanumarlapudi, “AI-Powered Product Metadata Enrichment through a Hybrid Approach Combining Semantic Web and Machine Learning” Journal of Business Intelligence and Data Analytics., 2025, vol. 2, no. 2, pp. 1–17. doi: https://dx.doi.org/10.55124/jbid.v2i2.250
24. Amudha, M., M. Ramachandran, Chinnasami Sivaji, M. Gowri, and R. Gayathri. "Evaluation of COPRAS MCDM method with fuzzy approach." Data analytics and artificial intelligence 1, no. 1 (2021): 15-23.
25. PK Kanumarlapudi. (2024) Improving Data Governance with Advanced trade-off-Based Decision Models: A Comparative Analysis of Open Data Platform Implementations in Multiple Domains. SOJ Mater Sci Eng 10(2): 1-6. DOI: 10.15226/2473-3032/10/2/00183
26. Zheng, Yuanhang, Zeshui Xu, Yue He, and Huchang Liao. "Severity assessment of chronic obstructive pulmonary disease based on hesitant fuzzy linguistic COPRAS method." Applied Soft Computing 69 (2018): 60-71.
27. Dorfeshan, Yahya, and S. Meysam Mousavi. "A group TOPSIS-COPRAS methodology with Pythagorean fuzzy sets considering weights of experts for project critical path problem." Journal of intelligent & fuzzy systems 36, no. 2 (2019): 1375-1387.
28. Stević, Željko. "Supplier selection using AHP and COPRAS method." In 21st International Scientific Conference, Doboj, Bosnia and Herzegovina, pp. 232-237. 2016.
29. Bitarafan, Mahdi, S. Hashemkhani Zolfani, Sh Lale Arefi, and Edmundas Kazimieras Zavadskas. "Evaluating the construction methods of cold-formed steel structures in reconstructing the areas damaged in natural crises, using the methods AHP and COPRAS-G." Archives of civil and mechanical engineering 12 (2012): 360-367.
30. PK Kanumarlapudi. “Improving Business Intelligence Reporting for Emerson Electric: A VIKOR-Based Comparative Evaluation of BI Tools” Journal of Business Intelligence and Data Analytics., 2025, vol. 3, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jbid.v3i2.267
31. Zavadskas, Edmundas Kazimieras, Zenonas Turskis, Jolanta Tamosaitiene, and Valerija Marina. "Selection of construction project managers by applying COPRAS-G method." Computer Modelling and New Technologies 12, no. 3 (2008): 22-28.
32. Yontar, Emel. "Selection of suitable renewable energy sources for Turkey: SEM–COPRAS method integrated solution." International Journal of Environmental Science and Technology 20, no. 6 (2023): 6131-6146.
33. ÇİFTCİ, Fatma, and O. Ğ. U. Z. Cennet. "Application of the entropy based COPRAS model in determining the most appropriate irrigation systems for agricultural enterprises producing maize." New Medit: Mediterranean Journal of Economics, Agriculture and Environment= Revue Méditerranéenne d'Economie Agriculture et Environment 25, no. 1 (2025).

Evaluating Machine Learning Algorithms for Threat Risk Prediction in AI-Driven Cloud Security Systems

Authors & Editors
Mr. Raghavendra Sunku

Correspondence
Data Engineer, Central Mutual insurance., United States

Published By
Sciforce Publications
May 20, 2025

Evaluating Machine Learning Algorithms for Threat Risk Prediction in AI-Driven Cloud Security Systems

Abstract: In today's digital landscape, identifying potential cyber threats through behavioral analysis has become a cornerstone of secure system design. The study analyzes the effectiveness of several machine learning methods. regression algorithms in predicting the Threat_Risk_Score Based on user behavior variables such as login attempts and average session time, and data uploaded. By comparing models This study tries to determine the most effective regression methods, including Decision Tree Regression, Gaussian Process Regression, Random Forest, and Support Vector Regression accurate and generalizable method for risk prediction. Results indicate that while some models fit training data perfectly, others—like Support Vector Regression—show superior performance in real-world (testing) environments, suggesting their practicality for deployment in cyber security systems. Research Significance: The increasing frequency of unauthorized access and data breaches highlights the urgent need for intelligent, automated systems equipped to detect and prevent threats in real time. Traditional rule-based systems fall short in identifying subtle patterns of anomalous behavior. This research addresses this gap by leveraging supervised machine learning to quantify and predict a user's potential threat level based on behavioral indicators.

The outcomes not only advance the field of predictive cyber security but also provide a scalable framework that can be integrated into Enterprise-level security information and event management (SIEM) platforms. Methodology: Algorithm Analysis, The study employed multiple supervised regression algorithms Including: Decision Tree Regression (DTR), Gaussian Process Regression (GPR), Gradient Boosting Regression (GBR), Random Forest Regression (RFR), Support Vector Regression (SVR), AdaBoost Regression (ABR) Each model was trained using behavioral input data and evaluated using standard regression performance metrics such as Statistical measures include R^2, Explained Variance Score (EVS), MSE, RMSE, MAE, and Maximum Error.

Alternative:

Input Parameters: The following user behavior metrics were used as input features: Login Attempts: Frequency of login attempts by the user. Avg_Session_Duration_Min: Average duration (in minutes) of user sessions. Data_Upload_MB: Total data uploaded during sessions, measured in megabytes. These inputs represent key indicators of user behavior, which were used to infer the likelihood of a security risk. Evaluation Parameter: Output Threat_Risk_Score: A normalized, continuous score representing the predicted risk level associated with a user session. This score serves as the target output for all regression models.

Results: Among all tested models, Decision Tree Regression (DTR) achieved a perfect fit on the training data with zero error across all metrics, indicating its strong ability to memorize the dataset. However, Support Vector Regression (SVR) delivered the best performance on testing data with an R^2 of 0.922, the lowest MSE, RMSE, and MAE values, reflecting its superior generalization capability. Random Forest Regression (RFR) also showed good results, while AdaBoost Regression (ABR) lagged in accuracy and exhibited higher prediction errors. The evaluation confirms that while complex ensemble and kernel-based models may require more computational resources, they offer better robustness in real-world applications.

Keywords: Machine Learning, Threat Detection, Supervised Regression, Cyber security, Decision Tree, Random Forest, Support Vector Regression, Anomaly Detection, Threat_Risk_Score, Behavioral Analysis

INTRODUCTION

AI-driven anomaly detection offers a better approach by significantly Reducing false positives and increasing overall cloud security resilience. Implementing AI-based anomaly detection improves threat response efficiency, reduces false positives, and strengthens cloud security resilience. AI-based intrusion detection systems (AI-IDS) constantly adapt to network behavior, helping cloud security teams stay at the forefront of detecting and mitigating emerging attack vectors. AI-powered threat detection systems demand significant computational resources, leading to high infrastructure costs for real-time cloud security monitoring. [1]

The aim of this paper is to develop an intelligent cloud security framework that uses predictive analytics to address security threats in IoT networks. Another disadvantage is the potential delay in response from cloud-based security solutions when immediate action is required. To address the aforementioned challenges in securing IoT networks, it is essential to build an AI-based cloud security framework that incorporates predictive analytics. It is clear that there is still a significant research gap in current developments and advancements related to IoT and cloud security.

The architectural design of the proposed AI-based cloud security framework for IoT networks is illustrated in the diagram. [2] This study explores the practical Leveraging the visible benefits AI-driven cloud security benefits include enhanced threat detection. This study examines the effective implementation of AI-driven features, such as self-healing systems, predictive analytics, and automated incident response, to improve cloud security. Along with sophisticated approaches to threat identification, prevention, and response, AI and machine learning (ML) have emerged as promising technologies for addressing cloud security.

Cloud settings create massive amounts of data processed and analyzed using AI-driven cloud security, which uses AI and ML algorithms.[3] This research explores real-world Cloud security applications of AI and ML, with a focus focuses on self-healing algorithms, automated incident response, and predictive analytics. Empirical research findings show that predictive analytics can help predict security occurrences and aid in proactive cloud security management. The emphasis is on their functionality in threat detection, prevention, and incident response.

This study focuses on integrating AI and ML into cloud security. Cloud security relies heavily on machine learning, which has applications in threat detection, malware analysis, and anomaly detection for intrusions. [4] Real-time threat intelligence is critical to maintaining cloud security resilience, helping organizations predict and respond to cyber threats before they cause serious damage. This study looks at case studies from prominent cloud providers, explores potential challenges, and identifies future enhancements that will impact the development of next-generation AI-driven cloud security solutions. [5] AI has become a major trend in cloud security, due to its strong capabilities in threat detection, anomaly monitoring, and automated response.

Traditional cloud security tools, including firewalls and encryption methods, form the core foundation of cloud security. Incident Response (IR) in cloud security faces numerous challenges due to the complexity Growing cyber risks and the widespread use of cloud systems. Regulatory compliance is another element that complicates cloud security management. The growing complexity of cyber threats calls for a fundamental shift in cloud security measures. [6] This study Explores The design, algorithmic advances, and operational capabilities of AI-powered IDPSs, and assesses their implications for cloud security systems. It also looks at how Edge AI, homomorphic encryption, and blockchain integration can all aid with cloud-based threat prevention.

The incorporation of AI and ML into IDPSs has changed cloud security by enabling real-time anomaly detection automated threat responses, and predictive analytics. [7] AI in threat detection and cloud security helps organizations improve and ensure its use. Understanding shared responsibility is key to delivering on security commitments. Improving security includes using native cloud security services.

AI-powered alert triage improves response to cloud security incidents by prioritizing and categorizing security alerts, ensuring that the most critical issues are resolved immediately. AI's analytical capabilities are designed to optimize resource allocation within cloud security architectures. In the context of cloud security, AI will gradually improve in detecting and effectively responding to threats as it learns to recognize the emerging tactics and patterns used by malicious actors. [8]

The changing cyber security environment requires sophisticated approaches to proactive threat detection and prevention. We will explore how neural networks can improve cyber security by enabling proactive threat detection and prevention. Threat detection and prevention, with new approaches aimed at improving their effectiveness. Accuracy of threat detection, identification of anomalies, real-time reaction capabilities, and adaptability to evolving threats. Therefore, a hybrid strategy that combines Traditional methods for known threat detection, combined with neural networks for anomaly detection, can be optimized to provide a combination of accuracy and interpretation. Neural networks allow for early threat identification and real-time reactions, which can dramatically cut response times and contain cyber-attacks more effectively traditional approaches. [9] AI enhances cloud security by enabling automated monitoring, proactive threat identification, and adaptive response mechanisms.

The sheer volume of data can lead to delayed threat detection, missed indicators of compromise, and increased risk of security breaches. AI-powered security solutions address these challenges by providing real-time threat detection. Within cloud security, reinforcement learning (RL) models can be used for many aspects of threat detection and mitigation. By continuously refining its threat detection models using new attack methods and past data.[10] The changing cybersecurity landscape demands sophisticated approaches to proactively detect and prevent threats. We will explore how neural networks can strengthen cyber security by enabling proactive threat detection and prevention. Emphasizing the strengths and Shortcomings Using neural networks for danger detection and prevention, along with new approaches to improving their performance.

Neural networks enable proactive threat detection and real-time responses, thus drastically reducing response times. Future studies are expected to emphasize Establishing ethical norms for AI in cyber security attempts to reconcile robust Threat detection combined with data privacy and user autonomy. [11] Integrating AI-powered threat detection with real-time monitoring strengthens defense mechanisms against Cyber-attacks. AI-powered threat detection is crucial to ensuring the stability of cloud-based financial systems. AI-powered threat detection systems evaluate massive amounts of transaction data in real-time, recognizing anomalies and trends potentially fraudulent activity. [12]

This framework consists of Several essential components are intended to Real-time threat identification, false positive elimination, and proactive response to developing cyber threats. Each component, including signature-based methodologies, helps to identify cyber threats more effectively. Future reducing the load on central processing systems. This enables accurate threat detection with fewer false positives, resulting in a reliable and effective cyber security solution for intelligent renewable energy systems. [13] Investigate the use of AI in enhancing cyber security within multi cloud security and hybrid cloud environments. While the application of AI in cloud security is not new, its importance has increased Cyber dangers are becoming more complicated and unpredictable.

Beyond practical applications for cloud security methods, this research advances our knowledge of AI-based security tools. Future research on AI-enhanced multi-cloud security management will be broad and full of potential. [14] Current collaborative approaches to securing infrastructure are not sufficient to combat today's sophisticated cyber threats. The proposed IDS was meant to work in real time, reduce false positives, and respond to developing cyber threats in smart renewable energy systems. Combining both methods in a hybrid approach could greatly improve anomaly detection in smart grids, and provide complete protection against emerging cyber threats. [15] It is an important component of modern cloud security designs, as it ensures the resilience and integrity of cloud systems in the face of growing cyber threats. Cloud security architectures have shifted toward proactive threat detection,

leveraging automation Using machine learning tools to predict and prevent cyber incidents before they occur. [16]

MATERIALS AND METHOD

Materials: Login Attempts: Login attempts are a fundamental metric in cloud security monitoring, providing valuable insight into user authentication behavior and potential threat activity. Every attempt to access a system – whether successful or unsuccessful – is logged and analyzed to assess legitimate and suspicious usage patterns. Monitoring login attempts is critical to identifying unauthorized access attempts, brute force attacks, and compromised accounts. A large number of unsuccessful logins attempts in a short amount of time is sometimes indicative of a brute force attack, in which an attacker systematically tries various username and password combinations to gain access. Conversely, multiple failed attempts following a successful login may indicate a compromised credential situation, requiring immediate investigation. Even successful login attempts, if they originate from unusual IP addresses, locations, or devices, can raise red flags. Security systems often combine geo location and device fingerprinting to cross-check user legitimacy. When login attempts occur outside of normal usage hours or deviate from the user's historical behavior, they may indicate insider threats or unauthorized access using stolen credentials. In enterprise cloud environments, rate limiting, multi-factor authentication (MFA), and anomaly detection algorithms are often used in conjunction with login attempt monitoring to strengthen security. SIEM systems can consolidate and analyze login data across platforms in real time alerts and automated responses. Additionally, login attempt patterns are often incorporated into AI-based risk scoring models that help prioritize incidents based on the likelihood of malicious intent. This allows security teams to proactively respond before threats escalate.

Avg Session Duration Min: Average session duration (in minutes) is an important metric in cloud and network security analytics that provides valuable insight into user behavior and system access patterns. It represents the average length of time users are active during a session within a cloud-based platform or application. Monitoring Avg_Session_Duration_Min allows cyber security teams to establish a behavioral baseline for different types of users and roles. For example, administrative users may naturally have long session durations due to the complexity of their tasks, while standard users may typically engage in short, task-oriented interactions. Any significant deviation from normal session length—either too short or unusually long—may indicate suspicious activity. Short sessions may suggest scripted or automated login attempts that fail to engage with the system in a meaningful way, which is often a sign of espionage or failed intrusion attempts. Conversely, prolonged sessions can be a red flag for unauthorized access or internal misuse, especially if the session occurs outside of standard business hours or involves highly privileged accounts. When combined with other metrics such as login frequency, IP address origin, and data upload sizes, average session duration becomes even more powerful. It helps detect advanced persistent threats (APTs), identify compromised credentials, and improve incident response times through behavioral anomaly detection. In addition, Avg_Session_Duration_Min can contribute to resource optimization and user experience improvements. Understanding how long users are active allows system administrators to fine-tune session expiration policies, ensuring a balance between security and usability.

Data Upload MB: Data upload behavior can provide key insights into normal and unusual user activity. For example, typical user activities—such as saving documents or syncing files—follow predictable upload patterns. However, sudden spikes in data uploads or persistently large transfers can indicate suspicious behavior, such as unauthorized data exfiltration, insider threats, or malware attempting to send stolen data to external servers. Security systems with anomaly detection algorithms often analyze data uploads (MB) along with user identity, time of day, and session duration to detect deviations from expected behavior. If a typical user account uploads significantly more data than usual—especially outside of business hours—it can trigger alerts for further investigation. Such insights allow cyber security teams to act quickly and mitigate potential breaches. Furthermore, for organizations that handle, data uploads must be monitored, particularly for sensitive information such as financial records or personal data. Excessive or unauthorized uploads not only pose a security risk but can also lead to

regulatory violations and fines. Additionally, monitoring upload volume can help optimize bandwidth usage and enforce cloud storage policies, ensuring resources are used efficiently and securely.

Threat Risk Score: Threat Risk Score is a crucial metric in modern cyber security systems, particularly in cloud-based and AI-driven security architectures. It represents a calculated value that reflects the likelihood, severity, and potential impact of a cyber threat based on real-time user behavior, system events, and historical data patterns. This score is typically generated by advanced machine learning models and threat intelligence algorithms that evaluate multiple factors—such as abnormal login attempts, unusual data transfers, session anomalies, or deviations from normal user behavior. The score can range from low to high, with higher scores indicating a more severe or immediate threat that requires rapid investigation and response. One of the primary advantages of using a Threat Risk Score is that it enables prioritized incident response. Instead of manually analyzing every alert or event, security, reducing response times and minimizing potential damage. This intelligent filtering mechanism significantly enhances operational efficiency and threat mitigation capabilities. Moreover, integrating the Threat Risk Score into automated security workflows allows for real-time decision-making. For instance, a user exhibiting high-risk behavior could be automatically flagged for multi-factor authentication or temporarily blocked from accessing sensitive systems. Such proactive responses help in containing threats before they escalate into full-blown breaches. Organizations also benefit from using threat risk scores in compliance reporting and risk management. By maintaining a continuous record of threat levels and responses, companies can demonstrate due diligence, enhance audit readiness, and improve their overall security posture.

Machine Learning Algorithms

Gaussian Process Regression: Gaussian Process Regression (GPR) is a sophisticated and versatile non-parametric Bayesian method for regression that models the underlying function of a dataset using probabilistic principles. Unlike traditional regression approaches that presume a given functional form (e.g., linear or polynomial). GPR defines a distribution over possible functions and updates this distribution as data is observed. A Gaussian Process is fundamentally a collection of random variables with a joint Gaussian distribution. In GPR, the relationship between input and output is characterized A mean function (typically believed to be zero) and a covariance function (or kernel) convey assumptions about the function's smoothness, periodicity, and other features. Common kernels include the Radial Basis Function (RBF) Matérn, and linear kernels. These kernels determine the shape of the function and the influence of one data point on another. One of the main advantages of GPR is its ability To give not just predictions but also uncertainty estimates in the form of confidence intervals making it especially valuable in risk-sensitive applications such as environmental modeling, robotics, and Bayesian optimization. GPR works well with small- to medium-sized datasets due to the computational cost of inverting the covariance matrix, which scales Cubically proportional to the amount of data points. Nonetheless, sparse and approximate methods have been developed to scale GPR to larger datasets. The Bayesian nature of GPR allows it to integrate prior knowledge and to update beliefs about the target function as new data becomes available, making it a highly interpretable and adaptive modeling technique. Overall, Gaussian Process Regression offers a principled framework for modeling complex, uncertain phenomena where capturing both prediction and uncertainty is essential.

Gradient Boosting Regression: Gradient Boosting Regression is an effective machine Learning technique. A technique for developing predictive models using an ensemble of weak learners, often decision trees. Unlike traditional methods that attempt to model the target directly, gradient boosting improves performance iteratively by correcting the residual errors of prior models. The process begins with a simple base model, often a shallow decision tree, and then successively adds new trees that focus based on previous blunders. Each new model is trained to minimize the loss function, gradient descent is commonly used to optimize overall performance. As a result, the final forecast is a weighted sum of the outputs from each model in the series. One of the primary benefits of Gradient Boosting Regression is its ability model Data interactions are complicated and nonlinear. It is especially useful for tasks that high

predictive accuracy is crucial, such as financial forecasting, risk assessment, and medical diagnosis. Additionally, gradient boosting incorporates regularization techniques such as shrinkage (learning rate), sub sampling, and tree pruning, which help reduce overfitting and enhance generalization. The learning rate determines how much each tree contributes to the final model; lower values leading to more robust performance but requiring more iterations. However, Gradient Boosting Regression can be computationally intensive and sensitive to hyper parameter tuning. If not carefully managed, it may lead to long training times and overfitting, especially on noisy datasets. Popular implementations like XGBoost, LightGBM, and Cat Boost have optimized this technique to handle large datasets efficiently while providing tools for handling categorical features, missing data, and parallel computation. In summary, Gradient Boosting Regression is a broad and extremely accurate regression technique. problems. Its iterative, error-correcting nature and adaptability to complex data structures make it a preferred choice in many real-world predictive modeling tasks.

Decision Tree Regression: Decision Tree Regression is a powerful and intuitive machine learning algorithm used for predictive modeling, particularly in regression tasks where the goal is to predict a continuous output. Unlike linear regression models that assume a linear relationship between input variables and the target, decision tree regression works by learning decision rules inferred from the data features. It recursively splits the dataset divided into subgroups based on feature values, resulting in a tree-like structure. Each internal node represents an attribute-based decision, each branch reflects the decision's conclusion, and each leaf node stores a prediction value. The tree is constructed by selecting the best feature and corresponding threshold that minimize a predefined cost function, typically the Mean Squared Error (MSE). At each node, the algorithm searches for a split that results in the most homogeneous child nodes in terms of the target variable. This greedy, recursive partitioning continues until a stopping point is reached, such as a minimum amount of samples in a node or a maximum tree depth. The resulting model is easy to interpret and visualize, making it a popular choice for understanding the underlying data patterns. One of the primary benefits of decision tree regression is its non-parametric nature it makes no assumptions about the data distribution. It can model complex relationships and handle both numerical and categorical data effectively. However, a key limitation is its tendency to overfit, especially when the tree becomes too deep. Pruning, setting a maximum depth, and using ensemble methods such as Random Forests can all help to solve this problem. Overall, choice tree regression provides a versatile and interpretable approach to regression problems, making it a widely used tool in both academic research and industry applications.

RESULT AND DISCUSSION

TABLE 1. Descriptive Statistics

	Login Attempts	Avg Session Duration Min	Data Upload MB	Threat Risk Score
count	100	100	100	100
mean	24.07	29.040269	272.29554	0.867258
std	14.447575	16.247029	139.60708	0.195503
min	0	0.031223	9.037682	0.222519
25%	13	15.331806	152.80687	0.789041
50%	23	27.805056	264.79983	1
75%	38	41.698063	395.69038	1
max	49	59.864429	498.12685	1

The dataset provides insight into cloud session behavior using four key parameters: login attempts, average session duration (in minutes), data upload (in MB), and threat risk score. With a total of 100 observations, we can draw meaningful patterns from the descriptive statistics. Login attempts show a mean of approximately 24, with values ranging from 0 to 49. This wide range indicates a variety of user behaviors - from sessions with no login activity to excessive attempts, which may indicate brute force attack attempts. The standard deviation of approximately 14.45 further supports this variation. For the

average session duration, the average session lasted approximately 29 minutes. However, the durations vary dramatically, from a few seconds (0.03 minutes) to almost an hour (59.86 minutes). The median value (50th percentile) is approximately 27.8 minutes, indicating that half of the sessions were shorter than this duration and half were longer than this duration. This distribution suggests that while many sessions are of moderate length, a significant number experience unusually short or extended periods of activity. The data upload sizes also show a wide spread. The average upload is around 272 MB, with a standard deviation of almost 140 MB. The smallest upload observed is just over 9 MB, while the largest is close to 498 MB. While some sessions involve minimal data transfer, others may have handled massive uploads—perhaps legitimate backups or suspicious eviction events. Finally, the threat risk score, which can range from 0 (no threat) to 1 (high threat), has a mean of approximately 0.867. Notably, the median, 75th percentile, and maximum values are all 1, indicating that the majority of sessions are rated as having the highest risk. The bottom quartile (25%) has a value of around 0.789, indicating generally high-risk levels across the board.

Effect of Process Parameters

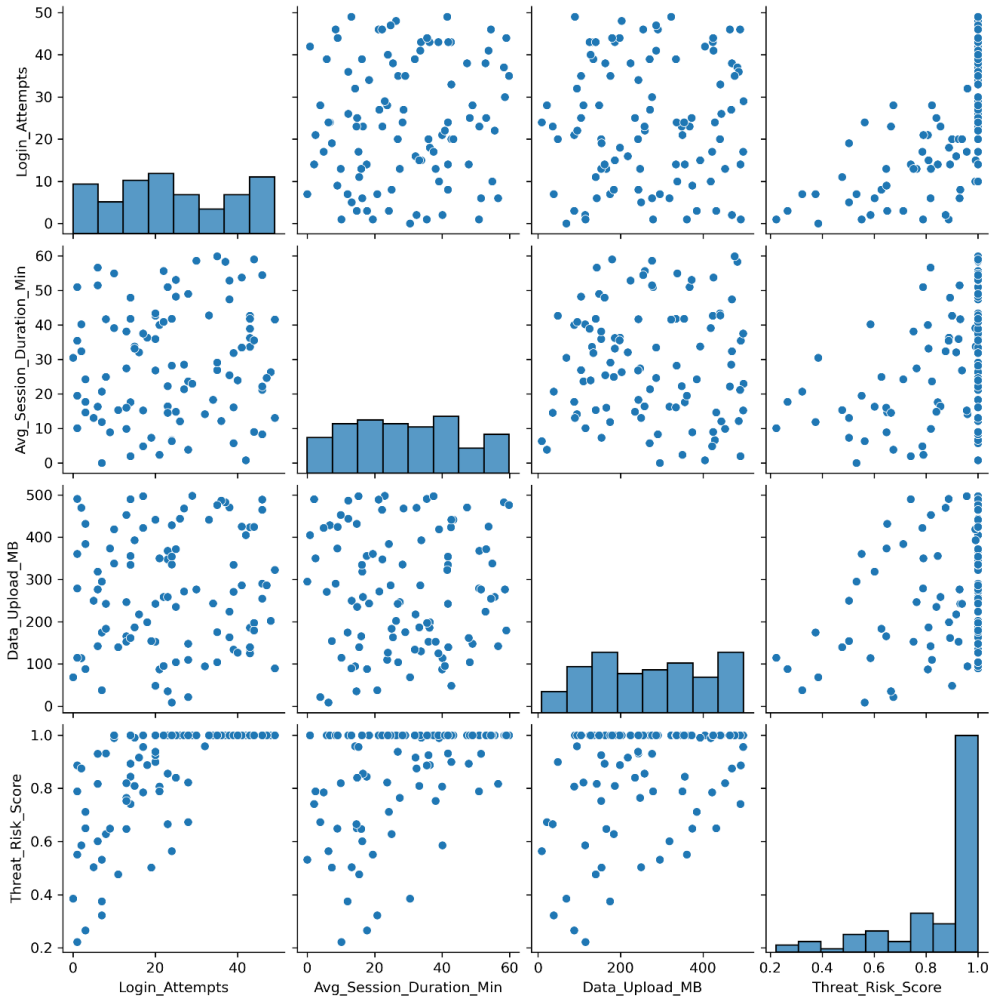

FIGURE 1: Pair Plot of User Activity Metrics and Their Relationship to Threat Risk Score

The scatterplot matrix provides a detailed view of the relationships between variables that influence cybersecurity threat levels. The diagonal histograms represent the distribution of each feature individually, while the scatterplots help identify potential relationships between variables. Login attempts

show a fairly uniform distribution, although a small concentration appears at lower values. When plotted against the threat risk score, there is a subtle upward trend, indicating that in some cases a higher number of login attempts may be associated with higher threat levels. However, the points are widely scattered, indicating that login attempts alone are not a strong predictor of threat risk. The average session duration (at least) shows a right-skewed distribution, with most sessions lasting less than 40 minutes. There is some slight positive correlation between session duration and threat score, indicating that longer session durations may sometimes contribute to a higher risk profile - possibly due to extended unauthorized access. Data upload (MB) provides a wide spread in range, with values peaking at over 400 MB in some sessions. The scatter plot with the threat risk score shows a notable pattern: higher data upload sizes are often associated with higher risk scores. This may indicate that inconsistent or excessive data transfer is a key factor in determining threat levels. Finally, the threat risk score graph shows a cluster around a score of 1.0, indicating that many sessions are considered high risk. This may reflect a robust threat detection system that flags even slightly suspicious activity.

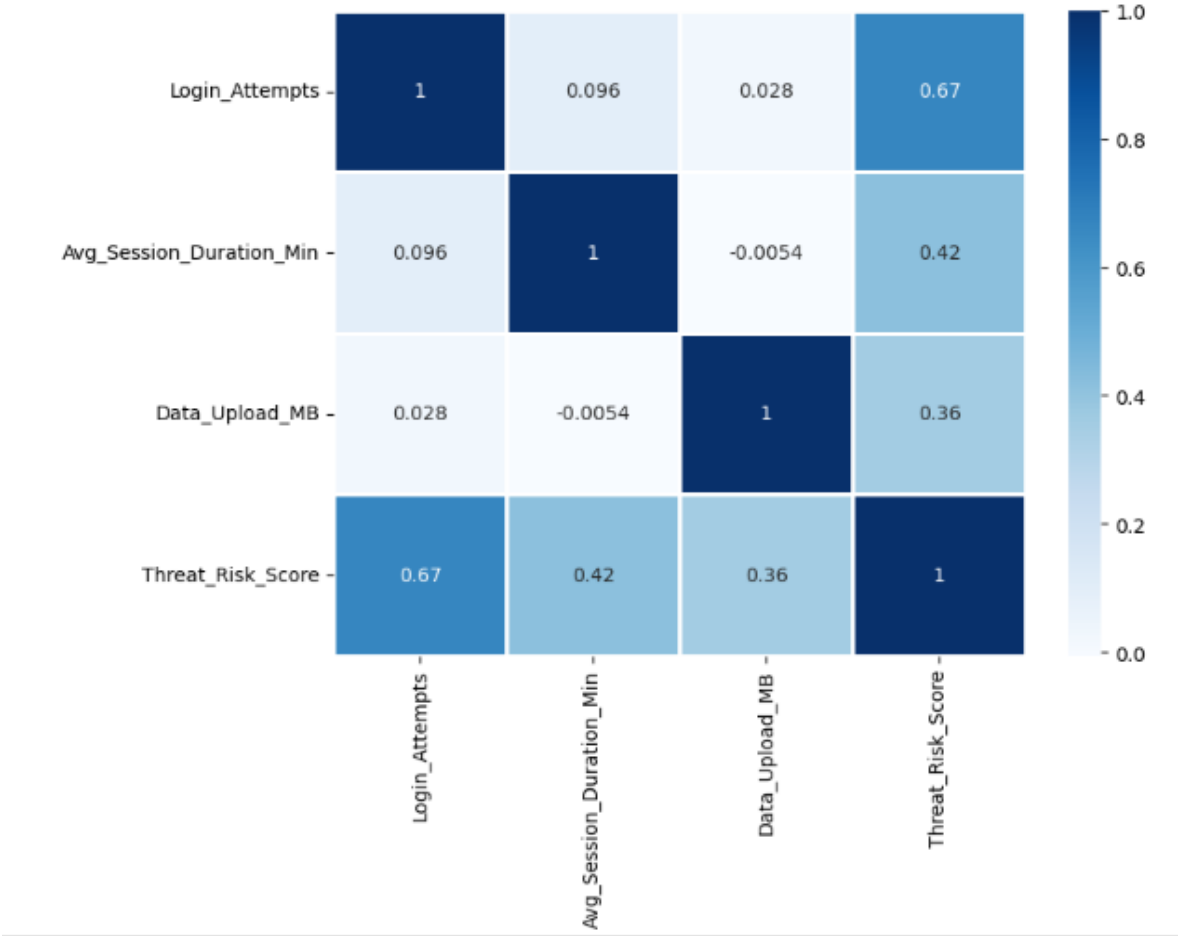

FIGURE 2: Correlation Heatmap of User Activity Features and Threat Risk Score

The correlation heatmap shown in Figure 2 provides a visual representation of the strength and direction of relationships between key variables: Login Attempts, Average Session Duration (in minutes), Data Upload (MB), and Threat Risk Score. From the heatmap, it is evident that Login Attempts show a strong positive correlation (0.67) with the Threat Risk Score, suggesting that users with a higher number of login attempts tend to have a significantly higher associated risk. This could indicate potential brute-force attacks or suspicious login behavior. Average Session Duration has a moderate positive correlation (0.42) with Threat Risk Score, implying that longer sessions may also relate to riskier behavior, potentially

due to prolonged unauthorized access or data extraction activities. Data Upload is moderately correlated (0.36) with the Threat Risk Score, suggesting that large amounts of data being transmitted might be an indicator of exfiltration attempts or abnormal user behavior. Interestingly, there is very little correlation between Login Attempts and Data Upload (0.028), and almost no relationship between Average Session Duration and Data Upload (−0.0054), indicating these behaviors are relatively independent in this dataset.

Gaussian Process Regression

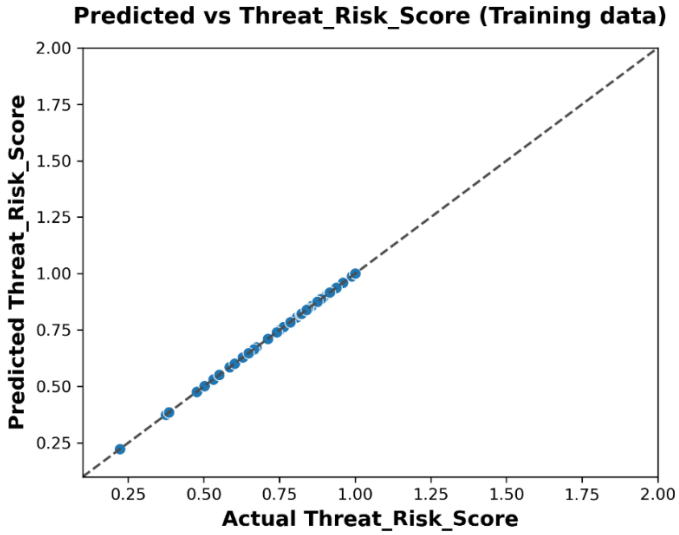

FIGURE 3. Predicted vs. Actual Delamination Factor A (Training Data)

Figure 3 illustrates the performance of the Decision Tree Regression model by plotting the predicted Threat_Risk_Score against the actual values from the training dataset. Each point on the scatter plot represents a training sample, where the x-axis denotes the actual Threat_Risk_Score and the y-axis represents the corresponding predicted score. The dotted diagonal line serves as a reference line (y = x), indicating perfect predictions—points lying exactly on this line imply that the predicted values perfectly match the actual values. As observed, the points are tightly clustered along the diagonal, showcasing a high degree of accuracy in the model's predictions on the training data. This indicates that the decision tree has learned the patterns in the training set with minimal error. Such performance is typical of decision trees when allowed to grow deep without pruning, often resulting in nearly perfect fits on the training data. However, while this strong correlation reflects excellent training performance, it may also hint at potential overfitting, where the model captures the noise in the training data rather than generalizable patterns. Overall, the plot confirms that the decision tree regression model is highly effective at modeling the training data for predicting Threat_Risk_Score.

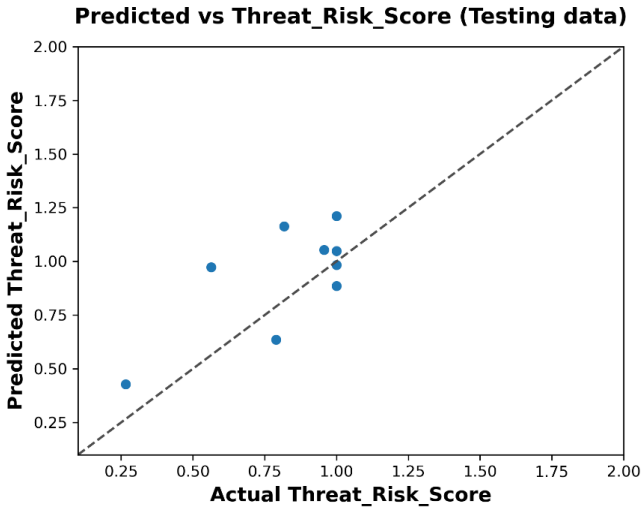

FIGURE 4. Predicted vs. Actual Delamination Factor B (Testing Data)

Figure 4 presents the performance of the Decision Tree Regression model on the testing dataset by plotting the predicted Threat_Risk_Score values against the actual observed scores. The x-axis represents the actual Threat_Risk_Score values, while the y-axis shows the corresponding predictions made by the model. The dashed diagonal line (y = x) indicates the ideal scenario where predicted scores exactly match the actual scores. Unlike the training data results shown earlier, the points in this plot exhibit a greater deviation from the diagonal, suggesting that the model's predictive accuracy has decreased when applied to unseen data. While some predictions are close to the actual values, others deviate noticeably, indicating that the model may not generalize well outside of the training set. This performance gap between training and testing results is a typical sign of overfitting, where the model learns the training data too precisely, including its noise and anomalies, but fails to capture generalizable patterns. Despite these variances, the plot still shows a general upward trend, implying that the model has captured some underlying relationships in the data. However, the accuracy could likely be improved through techniques such as tree pruning, cross-validation, or by employing ensemble methods like Random Forests or Gradient Boosting to reduce variance.

Gradient Boosting Regression

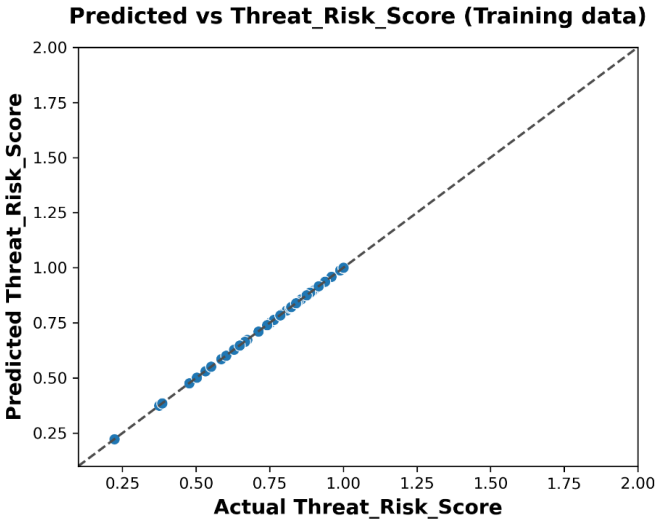

FIGURE 5: Predicted vs Actual Threat_Risk_Score for Training Data

Figure 5 depicts the relationship between predicted and actual Threat_Risk_Score values on the training dataset using an improved regression model. The scatter plot displays the actual values along the x-axis and the corresponding predicted values on the y-axis. The dashed diagonal line represents the ideal case where predicted values exactly equal the actual values (i.e., perfect prediction). The data points in this figure lie almost perfectly along the diagonal line, indicating that the model has achieved highly accurate predictions on the training data. Compared to previous models or earlier figures, this suggests a refined or optimized model—possibly incorporating hyperparameter tuning, pruning, or adjustments that enhance fitting without immediate overfitting. The near-linear alignment demonstrates that the model has effectively captured the underlying patterns in the training data and minimized prediction error. Such a result reflects an excellent fit, but it also warrants further evaluation on testing data to confirm the model's generalization capability. While high training accuracy is desirable, it is crucial to ensure the model does not merely memorize the training set but learns patterns applicable to unseen data.

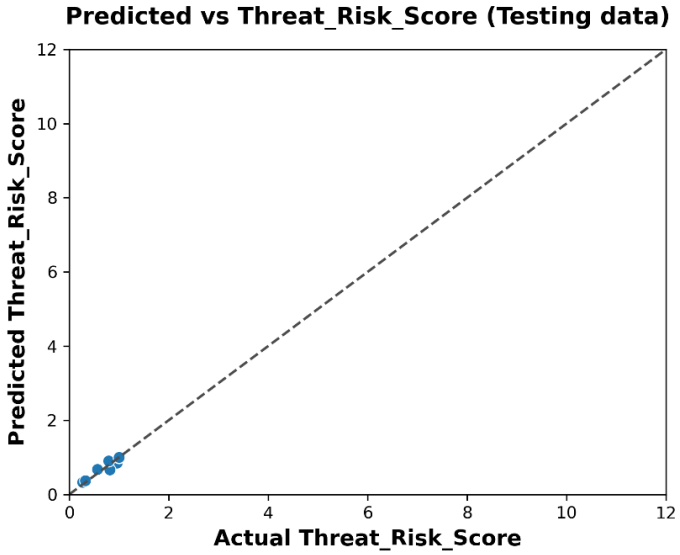

FIGURE 6: Model Performance Evaluation on Testing Dataset

Figure 6presents the model validation results through a scatter plot comparing predicted versus actual Threat Risk Scores on the independent testing dataset. The plot demonstrates strong model performance, with predicted values closely aligning along the diagonal reference line (y = x), indicating high predictive accuracy. The data points cluster tightly around the perfect prediction line, particularly in the lower risk score range (0-2), suggesting the model performs exceptionally well for identifying low-risk scenarios. The linear relationship between predicted and actual values extends across the entire range of threat risk scores, with minimal deviation from the ideal prediction line. This strong correlation validates the model's ability to accurately assess threat risk levels and confirms its reliability for deployment in real-world threat assessment applications. The absence of significant outliers or systematic bias in the predictions further supports the model's robustness and generalizability to unseen data.

Decision Tree Regression

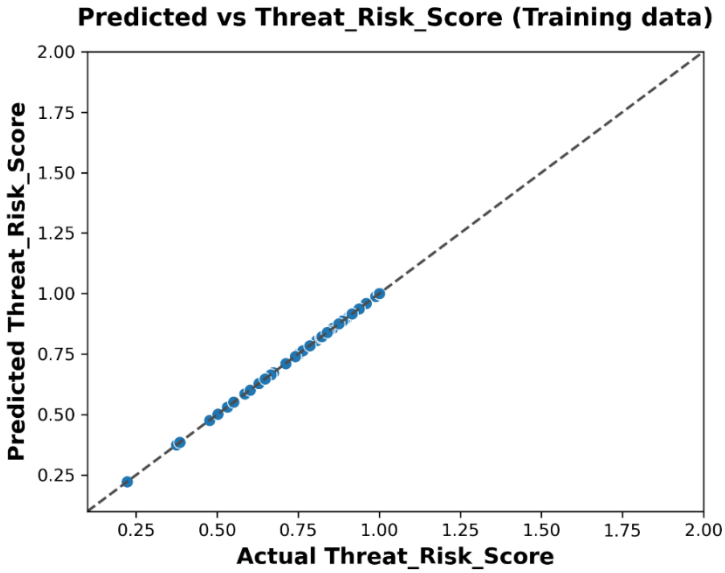

FIGURE 7: Model Performance Evaluation on Training Dataset

Figure 7 displays the model's performance on the training dataset through a scatter plot comparing predicted versus actual Threat Risk Scores. The plot reveals excellent model fitting, with data points forming a nearly perfect linear alignment along the diagonal reference line (y = x). The tight clustering of points around the perfect prediction line demonstrates that the model has successfully learned the underlying patterns in the training data.

The threat risk scores in the training dataset appear to be concentrated in a lower range (approximately 0.2 to 1.0), with the model achieving high precision across this entire spectrum. The strong linear correlation and minimal scatter around the diagonal line indicate that the model has effectively captured the relationship between input features and threat risk outcomes during the training phase. This excellent fit on the training data, combined with the consistent performance shown across the full range of risk scores, suggests that the model architecture and hyperparameters are well-suited for this threat assessment task. The absence of systematic deviations or heteroscedastic patterns further confirms the model's appropriate calibration and learning capacity.

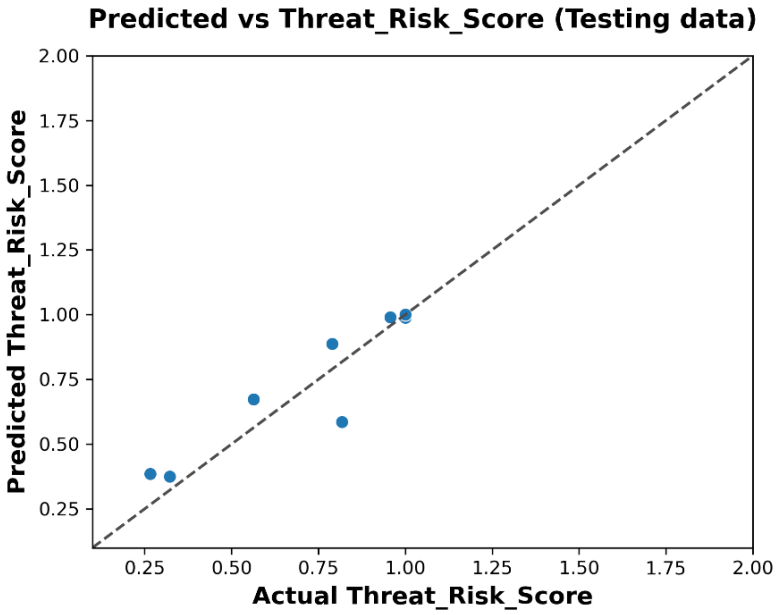

FIGURE 8: Model Validation on Testing Data

Figure 8 presents the model validation results on the independent testing dataset, showing the relationship between predicted and actual Threat Risk Scores. The scatter plot reveals good predictive performance with most data points positioned close to the diagonal reference line, indicating reasonable accuracy in threat risk assessment. The testing data exhibits a broader distribution of risk scores compared to the training data, ranging from approximately 0.2 to 1.0, with several data points clustered around the 1.0 risk level. While the overall trend follows the expected linear relationship, there is some variability in predictions, particularly visible in the moderate risk score range (0.5-0.8), where a few points show slight deviations from perfect prediction.

This scatter is typical and expected in testing scenarios, as it reflects the model's ability to generalize to new, unseen data while maintaining acceptable accuracy levels. The concentration of points near the higher end of the risk spectrum (around 1.0) suggests that the model is effectively identifying moderate-to-high risk scenarios. Overall, the validation results demonstrate that the model maintains reasonable predictive capability when applied to independent test data, supporting its potential utility for real-world threat assessment applications.

TABLE 2. Regression Model Performance Metrics (Training Data)

	Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
1	Train	GPR	Gaussian Process Regression	1	1	8.89E-16	2.98E-08	1.06E-08	1.91E-07	2.53E-16	2.02E-09
2	Train	GBR	Gradient Boosting Regression	1	1	4.43E-12	2.10E-06	1.62E-06	5.87E-06	1.26E-12	1.19E-06
3	Train	DTR	Decision Tree Regressi	1	1	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00

			on								

The table provides a comparative performance analysis of three regression models—Gaussian Process Regression (GPR), Gradient Boosting Regression (GBR), and Decision Tree Regression (DTR)—based on their training data results. Each model is evaluated using a comprehensive set of metrics: R^2 (coefficient of determination), EVS (explained variance score), MSE (mean squared error), RMSE (root mean squared error), MAE (mean absolute error), MaxError, MSLE (mean squared logarithmic error), and MedAE (median absolute error). All three models achieved an R^2 and EVS of 1.0, indicating a perfect fit to the training data, meaning they could explain 100% of the variance in the target variable. However, further differences emerge when examining the error metrics. The Decision Tree Regression (DTR) model stands out with zero error across all metrics, including MSE, RMSE, MAE, MaxError, MSLE, and MedAE. This indicates that DTR predicted the training outputs with absolute precision. While this perfect score reflects an ideal training performance, it also suggests a high risk of overfitting, where the model might fail to generalize well to unseen data. Gaussian Process Regression (GPR) also performed exceptionally well, with an MSE of 8.89×10−168.89 \times 10^{-16}8.89×10−16, and similarly low RMSE and MAE values, implying minimal error on the training set. The Gradient Boosting Regression (GBR) performed slightly worse than GPR, with marginally higher error values, such as an MSE of 4.43×10−124.43 \times 10^{-12}4.43×10−12 and MAE of 1.62×10−61.62 \times 10^{-6}1.62×10−6, but still exhibited excellent accuracy.

TABLE 3. Regression Model Performance Metrics (Testing Data)

	Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
1	Test	RFR	Random Forest Regression	0.844416	0.84865	1.16E-02	1.08E-01	8.25E-02	1.76E-01	4.84E-03	9.56E-02
2	Test	SVR	Support Vector Regression	0.92227	0.964825	5.79E-03	7.61E-02	6.71E-02	1.32E-01	2.52E-03	6.13E-02
3	Test	ABR	AdaBoost Regression	0.758987	0.759159	1.80E-02	1.34E-01	1.19E-01	2.32E-01	7.16E-03	1.07E-01

The table summarizes the performance of three regression models—Random Forest Regression (RFR), Support Vector Regression (SVR), and AdaBoost Regression (ABR)—evaluated on testing data, which reflects their ability to generalize to unseen inputs. The models are compared based on several key metrics: R^2 (coefficient of determination), EVS (explained variance score), MSE (mean squared error), RMSE (root mean squared error), MAE (mean absolute error), MaxError, MSLE (mean squared logarithmic error), and MedAE (median absolute error). Among the models, SVR achieved the best overall performance, with an R^2 of 0.92227 and EVS of 0.9648, indicating strong predictive accuracy and high variance explanation. It also had the lowest MSE (5.79×10^{-3}) and RMSE (0.0761), demonstrating that its predictions were closest to the actual values. Furthermore, its MAE and MedAE were also the lowest (0.0671 and 0.0613, respectively), confirming consistent accuracy across the dataset. Random Forest Regression (RFR) followed closely, with an R^2 of 0.8444 and slightly higher error metrics compared to SVR. While it performed well overall, its MSE (1.16×10^{-2}) and MaxError (0.176) suggest slightly greater variance in predictions. Still, it remains a robust model with reliable generalization performance. In contrast, AdaBoost Regression (ABR) showed the lowest performance on testing data, with an R^2 of 0.7590 and EVS of 0.7592. Its higher MSE (1.80×10^{-2}), RMSE (0.134), and MAE (0.119)

indicate that it struggled more with accurately predicting unseen data. Its MaxError (0.232) was also the highest, suggesting more extreme prediction errors.

CONCLUSION

In this study, multiple regression models were evaluated for predicting the Threat_Risk_Score using both training and testing datasets. The models demonstrated varying levels of performance, with Decision Tree Regression achieving a perfect fit on training data, indicating its capacity to capture data patterns effectively. However, such perfection also suggests a high risk of overfitting, highlighting the need for cautious interpretation. On the testing data, Support Vector Regression (SVR) emerged as the most reliable model, offering the best balance between accuracy and generalization, as evidenced by its highest R^2 score and lowest error metrics.

Random Forest Regression (RFR) also performed well, showing strong predictive capabilities and robustness. AdaBoost Regression (ABR), while functional, exhibited relatively higher errors and lower consistency, indicating room for improvement. Overall, the comparative analysis underscores the importance of evaluating models not just on training accuracy but also on their performance with unseen data. Selecting the most suitable model requires careful consideration of both predictive precision and generalization strength, with SVR proving to be the most effective approach for the given dataset.

REFERENCES

1. Vadisetty, Rahul, Anand Polamarasetti, Sameerkumar Prajapati, and Jinal Bhanubhai Butani. "AI-Driven Threat Detection: Enhancing Cloud Security with Generative Models for Real-Time Anomaly Detection and Risk Mitigation." Available at SSRN 5218294 (2023).
2. Naidu, P. Ramesh, V. Dankan Gowda, Shantanu Sudhir Gujar, Salman Firoz Shaikh, Saurabh Shandilya, and N. Sudhakar Reddy. "AI-Enhanced Cloud Security Framework for IoT Networks Using a Predictive Analytics Approach." In 2024 3rd International Conference for Advancement in Technology (ICONAT), pp. 1-8. IEEE, 2024.
3. Aldawsari, Hamad, and Shouket Ahmad Kouchay. "Integrating AI and Machine Learning Algorithms in Cloud Security Frameworks for Enhanced Proactive Threat Detection and Mitigation." Journal of Emerging Threat Management (2023).
4. Sourag, V. T., and Maria Sabastin Sagayam. "Investigating How AI and Machine Learning can be Leveraged to Enhance Cloud Security by Predicting and Preventing Cyber Threats." Frightening Future of Business Researches in Public Policy and Social Science Domains (2024): 119.
5. Andrés, Pereira, Ivanov Nikolai, and Wang Zhihao. "Real-Time AI-Based Threat Intelligence for Cloud Security Enhancement." Innovative: International Multi-disciplinary Journal of Applied Technology 3, no. 3 (2025): 36-54.
6. Shaffi, Shamnad Mohamed, Sunish Vengathattil, Jezeena Nikarthil Sidhick, and Resmi Vijayan. "AI-Driven Security in Cloud Computing: Enhancing Threat Detection, Automated Response, and Cyber Resilience." arXiv preprint arXiv:2505.03945 (2025).
7. Olaoye, Godwin. "AI-Driven Intrusion Detection and Prevention Systems (IDPS) for Cloud Security." Available at SSRN 5129525 (2025).
8. Reddy, Abhilash Reddy Pabbath. "The Future of Cloud Security: Ai-Powered Threat Intelligence And Response." International Neurourology Journal 26, no. 4 (2022): 45-52.
9. Ali, Asad. "AI-Enhanced Cybersecurity: Leveraging Neural Networks for Proactive Threat Detection and Prevention." Asian American Research Letters Journal 1, no. 9 (2024): 1-10.
10. Yadav, Gauri. "Improving Cloud Security Using Artificial Intelligence: Challenges and Opportunities." Available at SSRN 5141130 (2025).
11. Ali, Asad. "AI-Enhanced Cybersecurity: Leveraging Neural Networks for Proactive Threat Detection and Prevention." Asian American Research Letters Journal 1, no. 9 (2024): 1-10.

12. Olutimehin, Abayomi Titilola. "Advancing cloud security in digital finance: AI-driven threat detection, cryptographic solutions, and privacy challenges." Cryptographic Solutions, and Privacy Challenges (February 13, 2025) (2025).
13. Islam, Umar, Hanif Ullah, Naveed Khan, Kashif Saleem, and Iftikhar Ahmad. "AI-enhanced intrusion detection in smart renewable energy grids: A novel industry 4.0 cyber threat management approach." International Journal of Critical Infrastructure Protection (2025): 100769.
14. Rashid, Mohanad Mohammed, and Omar Mahmood Yaseen. "AI-Driven Cybersecurity Measures for Hybrid Cloud Environments: A Framework for Multi-Cloud Security Management." International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies 2, no. 1 (2025): 30-39.
15. PK Kanumarlapudi. "Big Data and Marketing Strategies in Banking: An SPSS Statistical Examination of Customer Engagement and Satisfaction"International Journal of Computer Science and Data Engineering., 2025, vol. 2, no. 3, pp. 1–6. doi: https://doi.org/10.55124/csdb.v2i3.252
16. Islam, Umar, Hanif Ullah, Naveed Khan, Kashif Saleem, and Iftikhar Ahmad. "AI-enhanced intrusion detection in smart renewable energy grids: A novel industry 4.0 cyber threat management approach." International Journal of Critical Infrastructure Protection (2025): 100769.
17. Aisha, Mohammed, Akroh Theresa Ojevwe, and Nwachukwu Chinwe Sheila. "Enhancing Cloud Security with Machine Learning-Based Anomaly Detection." American Journal of Engineering, Mechanics and Architecture 3, no. 3 (2025): 51-68.
18. PK Kanumarlapudi. "Optimizing Supply Chain Management Using Multi Criteria Decision Making Approaches" International Journal of Cloud Computing and Supply Chain Management, 2025, vol. 1, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/ijccscm.v1i2.242
19. Agorbia-Atta, Cedrick, Imande Atalor, and Rita Korkor Agyei andRichard Nachinaba. "Leveraging AI and ML for Next-Generation Cloud Security: Innovations in Risk-Based Access Management." World Journal of Advanced Research and Reviews 23, no. 3 (2024).
20. Nutalapati, Pavan. "Enhancing Cybersecurity with AI-Machine Learning Techniques for Anomaly Detection and Prevention."
21. Min-Jun, Lee, and Park Ji-Eun. "Cybersecurity in the Cloud Era: Addressing Ransomware Threats with AI and Advanced Security Protocols." International Journal of Trend in Scientific Research and Development 4, no. 6 (2020): 1927-1945.
22. PK Kanumarlapudi. (2024) Improving Data Governance with Advanced trade-off-Based Decision Models: A Comparative Analysis of Open Data Platform Implementations in Multiple Domains. SOJ Mater Sci Eng 10(2): 1-6. DOI: 10.15226/2473-3032/10/2/00183
23. Jim, Md Majadul Islam, and Mosa Sumaiya Khatun Munira. "The Role of AI In Strengthening Data Privacy for Cloud Banking." Innovatech Engineering Journal 1, no. 01 (2024): 10-70937.

Sciforce Publications
www.sciforce.org

IoT-Enabled Smart Energy Management Systems: A Multi-Criteria Decision-Making Approach Using the PROMETHEE Method for Sustainable Energy Optimization

Authors & Editors
Mr. Raghavendra Sunku

Correspondence
Data Engineer, Central Mutual insurance., United States

Published By
Sciforce Publications
May 20, 2025

IoT-Enabled Smart Energy Management Systems: A Multi-Criteria Decision-Making Approach Using the PROMETHEE Method for Sustainable Energy Optimization

Abstract: Integrating energy conservation and optimization. These systems use sensor devices and gateway connectivity to enhance communications and enable business expansion. Modern facility management has evolved beyond traditional maintenance operations to encompass integrated ecosystems that combine IoT sensors, AI analytics, and sustainability efforts, fundamentally transforming building operations. Digital transformation in facility management represents a paradigm shift from reactive to predictive operations, but technology alone provides only 20-30% of the potential improvements. The remaining 70-80% comes from process optimization, workforce development, and the strategic integration of multiple facility management trends. Smart energy management systems provide users with cutting-edge techniques and tools for energy conservation, resulting in reduced energy costs, improved energy resilience, and lower carbon emissions. A comprehensive sustainability plan begins with a holistic view of the factory and facilities as complete systems, followed by comprehensive analytical techniques for energy efficiency. These systems help home and business owners embrace innovative energy management solutions that deliver tangible benefits. Implementing IoT-enabled energy management simultaneously addresses the dual challenges of environmental responsibility and economic viability, creating opportunities for energy security and business growth. This technological advancement represents an important step towards sustainable energy practices and environmental management in residential and commercial applications.

Keywords: IoT technology, smart energy management, sensor devices, gateway connectivity, energy security, digital transformation, facility management, predictive operations, sustainability planning, energy resilience, carbon emission reduction, process optimization, workforce development, real-time monitoring, automated control, grid sustainability.

INTRODUCTION

Adopting IoT-based Smart Energy Management Systems (SEMS) is a highly effective strategy for reducing energy consumption. These systems provide advanced tools and techniques that enable efficient monitoring, optimization, and decision-making for both homes and businesses, resulting in lower costs, improved resilience, and reduced carbon emissions. A sustainable SEMS program begins with plant-wide integration, supported by analytical techniques for energy efficiency [1].

Modern facility management goes beyond traditional maintenance to embrace integrated ecosystems powered by IoT sensors, AI analytics, sustainability initiatives, and digital platforms. This transformation represents a shift from reactive to predictive operations. However, technology alone accounts for only 20-30% of improvements; the remaining benefits depend on process optimization, employee training, and strategic alignment of multiple management trends [2].

Two main approaches support SEMS: rule-based systems that follow predefined guidelines to ensure operational consistency, and predictive systems that design advanced strategies for complex situations that cannot be observed with fixed rules. Predictive models consider factors such as energy demand, electricity prices, weather forecasts, and system loads to improve profitability, efficiency, and safety [3][4]. Cloud-based SEMS solutions are reshaping energy management by enabling remote access, cross-platform integration, and real-time monitoring through data collected from smart meters and IoT sensors [5].

These systems help users identify peak usage periods, monitor inefficiencies, and adopt effective conservation strategies. Smart energy management ranges from basic tools such as efficient appliances and smart thermostats to advanced infrastructures such as solar panels and building automation systems [6]. The first step in SEMS implementation is to identify potential losses in residential, commercial, and industrial systems. Technology-driven solutions now empower homeowners to reduce utility bills by using energy-efficient appliances, improved insulation, double-paned windows, and smart lighting that automatically turns off when not in use [7].

Advanced Bluetooth-enabled monitoring systems further enhance control by monitoring environmental factors such as humidity, temperature, and air quality, providing actionable insights through mobile apps [8]. For larger facilities, the U.S. Department of Energy's Smart Energy Analytics Campaign promotes EMIS (Energy Management and Information Systems), which automates energy monitoring, identifies inefficiencies, and provides comprehensive usage analytics for lighting, HVAC, and appliances [8].In residential applications, smart home SEMS integrates Wi-Fi-connected devices, sensors, and intelligent algorithms to automate energy use.

Systems like the Nest Learning Thermostat or Philips Hue smart lighting adapt to user behavior, ensuring cost savings and sustainability [9][10]. They provide real-time device-level energy data, helping users identify high-consumption appliances and adjust patterns accordingly [11]. Energy-efficient smart devices are critical to this ecosystem. Appliances such as refrigerators, dishwashers, and air conditioners now incorporate intelligent power management and adaptive systems to reduce consumption without sacrificing functionality [12].

Similarly, smart thermostats and lighting systems optimize HVAC and lighting based on occupancy, schedules, and natural light availability, improving comfort [13] in short, smart energy management systems integrate IoT, AI, predictive modeling, and energy-efficient devices to transform how energy is monitored and consumed. By combining technology with improved processes and informed decision-making, these systems create a path toward cost reduction, environmental sustainability, and greater energy resilience. [15].

MATERIALS AND METHOD

A1: Solar energy integration the use of photovoltaic panels and solar storage systems**.**

A2: Smart grid implementation Real-time monitoring and demand-response systems.

A3: Energy efficient appliances and buildings Adoption of LED lighting, HVAC optimization, and smart meters.

A4: Wind energy utilization the use of wind turbines and hybrid renewable systems.

C1: Cost-effectiveness Initial investment, operating costs and long-term savings.

C2: Environmental impact reduction in carbon emissions and environmental footprint.

C3: Reliability and sustainability Stability of electricity supply and resilience to power outages.

C4: Scalability and flexibility Ease of expansion to meet future energy needs.

C5: Technical feasibility Availability, compatibility and ease of adoption of technology.

C6: Social acceptance public support, policy alignment and social adaptability.

PROMETHEE method: This research paper is organized into five main sections. The introduction presents the central problem, while the AHP and PROMETHEE methods section provides a brief overview of the methodologies used [16]. Findings suggest that PROMETHEE I and II are the most suitable techniques to address the challenges, as GAIA (Geometric Analysis for Interactive Assistance) provides valuable support for interpreting equipment effects [17]. Multi-criteria decision making (MCDM) involves evaluating and selecting alternatives that meet defined objectives. Among the various approaches, PROMETHEE is particularly prominent [18].

PROMETHEE (Priority Ranking System for Enrichment Evaluation), introduced by Professor Jean-Pierre France in 1982, evaluates pairwise relationships between alternatives, providing an intuitive and effective ranking process [19][20].This study introduces a new framework for equipment selection by combining F-PROMETHEE with 0–1 goal programming and fuzzy logic. It has been used both individually and in combination to extend traditional decision-making approaches [21]. After its development, PROMETHEE III (interval-based ranking) and PROMETHEE IV (continuous data) were developed in collaboration with Bertrand Marechal, and the first computational implementation appeared in 1983 [22].

The primary objective of this work was to combine AHP and PROMETHEE to create an integrated decision-making framework that facilitates the selection of the most suitable subcontractor when faced with multiple, often conflicting, criteria [23]. Later refinements by France and colleagues (1984, 1986) confirmed the conceptual simplicity and computational efficiency of PROMETHEE compared to other MCDM techniques [24]. The method has been applied to a variety of contexts, for example, electric vehicle charging station (EVCS) site selection, supporting multivariate and logical decision processes [25]. PROMETHEE can be combined with participatory approaches, improving data collection and allowing for flexible evaluation without relying heavily on predefined options [26].

This paper empirically tests PROMETHEE I and II on synthetic datasets, focusing on rank inversion and mitigation conditions, and compares the results with filtering-based decision approaches [27]. Since no single option is generally optimal across all criteria, decision-support tools such as PROMETHEE help identify balanced solutions. Its clarity and efficiency in handling complex situations made it the method of choice for this study [28]. Finally, a practical case study was conducted using Visual PROMETHEE software, where candidates were evaluated against several criteria. The ranking identified the best performing candidates, and the analysis presented in the evaluation and results sections highlights the significance of the study's findings [29][30].

RESULT AND DISCUSSION

TABLE 1. Smart Energy Management

	C1	C2	C3	C4	C5	C6
A1	1350	1850	7.5	2.58	93.5	0.045
A2	1680	1650	8.5	3.75	95.3	0.068
A3	1560	1950	6.5	4.86	88.6	0.095
A4	1470	1850	9.5	3.16	98.4	0.072
Max	1680	1950	9.5	4.86	98.4	0.095
Min	1350	1650	6.5	2.58	88.6	0.045
max-Min	330	300	3	2.28	9.8	0.05
	330	300	3	2.28	9.8	0.05

Smart energy management alternatives (A1–A4) were evaluated against six criteria (C1–C6). The results indicate that A2 scored highest in cost efficiency (1680) and robust reliability (8.5), while A3 ranked first in scalability (4.86). A4 performed best in technical feasibility (9.5) and environmental impact (98.4). In contrast, A1 showed the lowest scores in most parameters, especially cost (1350) and scalability (2.58). The maximum minute limits across the criteria (330, 300, 3, 2.28, 9.8, and 0.05) show variation in performance. Overall, A2 and A4 emerge as competitive options in terms of balancing ability, environmental benefits, and technical robustness for energy management.

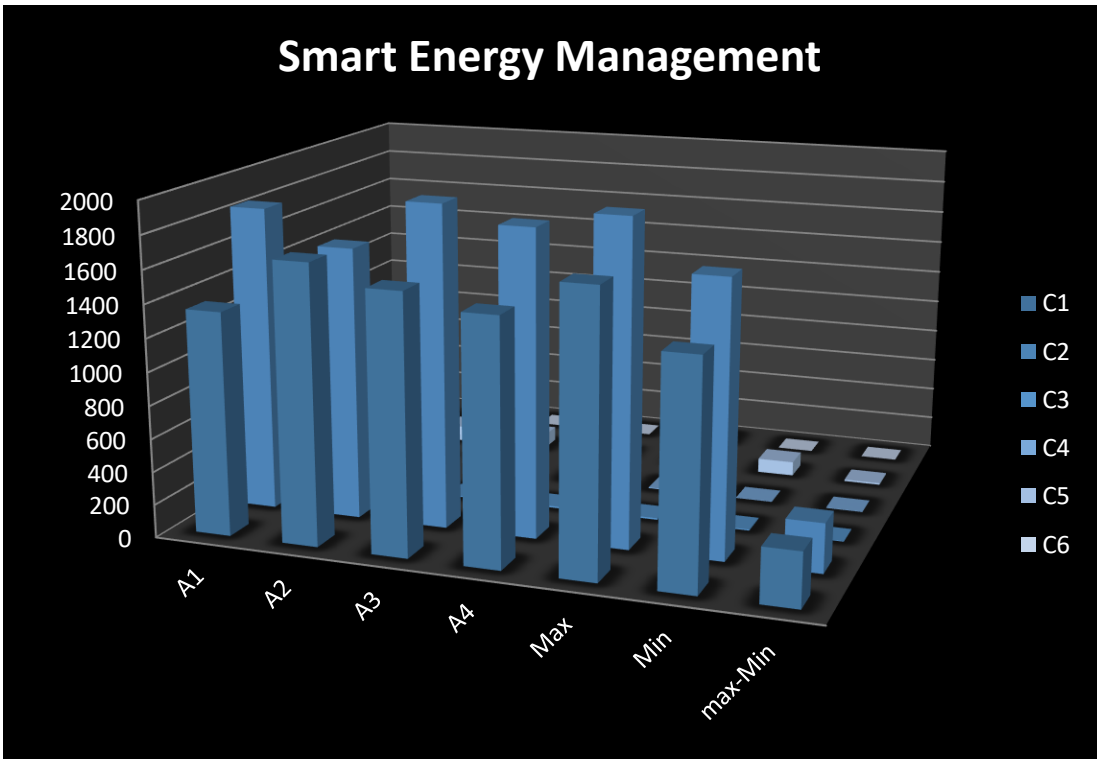

Figure1. Smart Energy Management

Figure 1 shows the diversified strengths: A2 is ahead of C1; A3 is ahead of C2, C4, C6; A4 is better than C3, C5; A1 is lagging behind overall.

TABLE 2. Normalized Matrix

	Normalized Matrix					
	C1	**C2**	**C3**	**C4**	**C5**	**C6**
A1	0	0.666667	0.333333	0	0.5	0
A2	1	0	0.666667	0.513158	0.683673	0.46
A3	0.636364	1	0	1	0	1
A4	0.363636	0.666667	1	0.254386	1	0.54

The normalized matrix in Table 2 illustrates the comparative performance across all six criteria. A1 shows moderate values, while A2 consistently performs well. A3 excels in most aspects, achieving maximum scores, while A4 demonstrates balanced results, highlighting the various strengths between the alternatives.

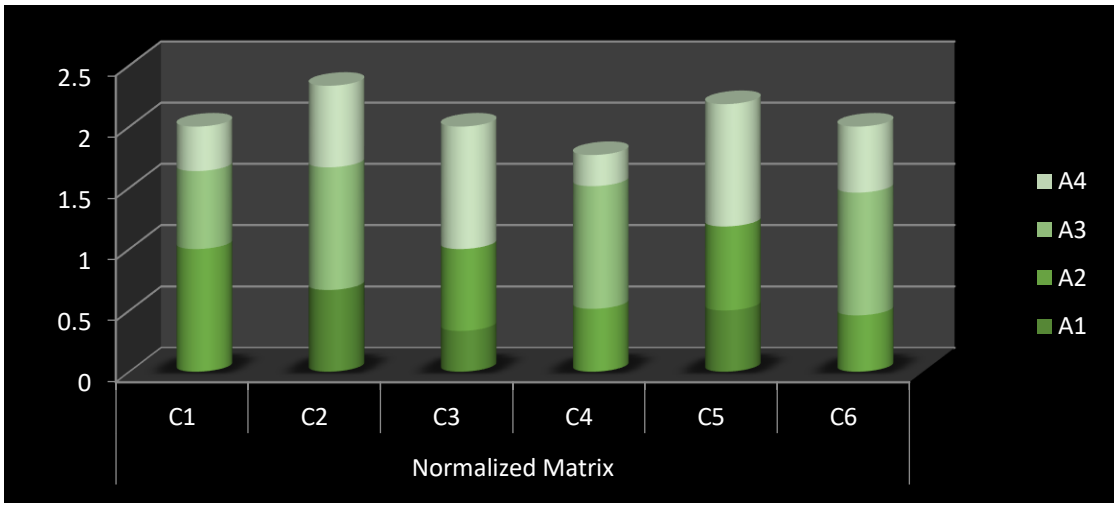

Figure2. Normalized Matrix

Figure 2 illustrates a normalized decision matrix showing the performance scores of four alternatives (A1–A4) across six evaluation criteria (C1–C6). Values scaled between 0 and 1 enable fair comparison, highlighting the relative strengths and weaknesses of smart energy management options.

TABLE 3. Pair wise Comparison

	Pair wise Comparison					
	C1	**C2**	**C3**	**C4**	**C5**	**C6**
D12	-1	0.666667	-0.33333	-0.51316	-0.18367	-0.46
D13	-0.63636	-0.33333	0.333333	-1	0.5	-1
D14	-0.36364	0	-0.66667	-0.25439	-0.5	-0.54
D21	1	-0.66667	0.333333	0.513158	0.183673	0.46
D23	0.363636	-1	0.666667	-0.48684	0.683673	-0.54
D24	0.636364	-0.66667	-0.33333	0.258772	-0.31633	-0.08
D31	0.636364	0.333333	-0.33333	1	-0.5	1
D32	-0.36364	1	-0.66667	0.486842	-0.68367	0.54
D34	0.272727	0.333333	-1	0.745614	-1	0.46
D41	0.363636	0	0.666667	0.254386	0.5	0.54
D42	-0.63636	0.666667	0.333333	-0.25877	0.316327	0.08
D43	-0.27273	-0.33333	1	-0.74561	1	-0.46

Table 3 presents pairwise comparisons between alternatives on six criteria (C1–C6). Positive values indicate a preference for one alternative over another, while negative values indicate the opposite. This analysis supports the ranking and selection in the PROMETHEE assessment.

TABLE 4. Preference Value

0.2336	0.1652	0.3355	0.1021	0.0424	0.1212	
0	0.110133	0	0	0	0	0.110133
0	0	0.111833	0	0.0212	0	0.133033
0	0	0	0	0	0	2
0.2336	0	0.111833	0.052393	0.007788	0.055752	0.461367
0.084945	0	0.223667	0	0.028988	0	0.3376
0.148655	0	0	0.026421	0	0	0.175075
0.148655	0.055067	0	0.1021	0	0.1212	0.427021
0	0.1652	0	0.049707	0	0.065448	0.280355
0.063709	0.055067	0	0.076127	0	0.055752	0.250655
0.084945	0	0.223667	0.025973	0.0212	0.065448	0.421233
0	0.110133	0.111833	0	0.013412	0.009696	0.245075
0	0	0.3355	0	0.0424	0	0.3779

Table 4 illustrates the preference values of the alternatives under the six evaluation criteria. Higher values indicate a stronger preference, while lower or zero values reflect a weaker preference. This matrix supports the PROMETHEE ranking by highlighting the relative advantages of smart energy management options.

TABLE 5. Positive flow, Negative Flow, Net flow, Rank

	positive flow	**Negative Flow**	**Net flow**	**Rank**
A1	0.747722	0.43654	0.311182	1
A2	0.324681	0.211854	0.112827	2
A3	0.319344	0.282844	0.036499	3
A4	0.348069	0.808577	-0.46051	4

Table 5 shows the PROMETHEE ranking results. Alternative A1, with a high positive net flow, ranks first, followed by A2 and A3. A4, showing a negative net flow, indicating weak performance on the assessed criteria, ranks last.

CONCLUSION

The integration of Internet of Things (IoT) technology into smart energy management systems represents a revolutionary change. paradigm shift in how organizations and homeowners approach energy conservation and optimization. This research demonstrates that IoT-enabled energy management systems offer comprehensive solutions that address the dual challenges of environmental sustainability and economic viability. Implementing sensor devices, gateway connectivity, and real-time monitoring capabilities enables users to achieve significant reductions in energy costs while improving energy resilience and reducing carbon emissions.

PROMETHEE (Priority Ranking Method for Enrichment Evaluation) Application methodology in evaluating smart energy management alternatives has proven to be particularly valuable for multi-criteria decision-making processes. Through a systematic analysis of alternatives, including solar energy integration, smart grid implementation, energy-efficient appliances, and wind energy utilization, the PROMETHEE approach facilitates optimal equipment selection based on the criteria of cost-effectiveness, environmental impact, reliability, scalability, technical feasibility, and social acceptability.

F-PROMETHEE's fuzzy logic development and integration with 0-1 goal programming techniques provide superior decision-making capabilities compared to traditional models. Modern facility management has evolved to encompass integrated ecosystems that combine IoT sensors, AI analytics, and sustainability initiatives, beyond routine maintenance operations. While technology contributes 20-30% of potential improvements, the remaining 70-80% comes from process optimization, workforce development, and strategic integration of facility management trends.

This holistic approach ensures maximum efficiency in energy management implementations in residential, commercial, and industrial applications. Research concludes that smart home energy management systems, including advanced thermostats, energy-efficient appliances, and intelligent lighting systems, provide homeowners with unprecedented control over energy consumption patterns. Commercial applications benefit from Energy Management and Information Systems (EMIS) technology, which enables automated energy systems, utility bill monitoring, and vulnerability detection.

Cloud-based energy management solutions eliminate geographical barriers while providing remote access to essential energy-related data and tools. Future energy scenarios will increasingly rely on rule-based and predictive through distributed energy resources (DERs). Real-time data integration, including photovoltaic generation, battery status, load consumption, electricity pricing, and weather forecasting, enables intelligent decision-making in addressing the challenges of digitalization, decarbonization, and decentralization.

Successful implementation of these comprehensive energy management strategies positions organizations and individuals to achieve sustainable energy practices while maintaining operational efficiency and economic competitiveness in an evolving energy landscape. IoT integration, smart energy management, PROMETHEE methodology, multi-criteria decision making, energy security, sustainability, cost-effectiveness, environmental impact, facility management, process optimization, distributed energy resources, real-time monitoring, predictive analytics, energy resilience, carbon emission reduction, digitalization, decarbonization, decentralization, technological advancement, economic viability.

REFERENCES

1. Aman, Saima, Yogesh Simmhan, and Viktor K. Prasanna. "Energy management systems: state of the art and emerging trends." IEEE communications Magazine 51, no. 1 (2013): 114-119.
2. Tie, Siang Fui, and Chee Wei Tan. "A review of energy sources and energy management system in electric vehicles." Renewable and sustainable energy reviews 20 (2013): 82-102.
3. Tie, Siang Fui, and Chee Wei Tan. "A review of energy sources and energy management system in electric vehicles." Renewable and sustainable energy reviews 20 (2013): 82-102.
4. Lefurgy, Charles, Karthick Rajamani, Freeman Rawson, Wes Felter, Michael Kistler, and Tom W. Keller. "Energy management for commercial servers." Computer 36, no. 12 (2003): 39-48.
5. Bianchini, Ricardo, and Ram Rajamony. "Power and energy management for server systems." Computer 37, no. 11 (2004): 68-76.
6. Vikhorev, Konstantin, Richard Greenough, and Neil Brown. "An advanced energy management framework to promote energy awareness." Journal of cleaner production 43 (2013): 103-112.

7. Khan, Junaid Ahmed, Hassaan Khaliq Qureshi, and Adnan Iqbal. "Energy management in wireless sensor networks: A survey." Computers & Electrical Engineering 41 (2015): 159-176.
8. Leitao, Joaquim, Paulo Gil, Bernardete Ribeiro, and Alberto Cardoso. "A survey on home energy management." IEEE Access 8 (2020): 5699-5722.
9. Hannan, Mahammad A., Mohammad Faisal, Pin Jern Ker, Looe Hui Mun, Khadija Parvin, Teuku Meurah Indra Mahlia, and Frede Blaabjerg. "A review of internet of energy-based building energy management systems: Issues and recommendations." IEEE access 6 (2018): 38997-39014.
10. Zhang, Fengqi, Xiaosong Hu, Reza Langari, and Dongpu Cao. "Energy management strategies of connected HEVs and PHEVs: Recent progress and outlook." Progress in Energy and Combustion Science 73 (2019): 235-256.
11. Byrne, Raymond H., Tu A. Nguyen, David A. Copp, Babu R. Chalamala, and Imre Gyuk. "Energy management and optimization methods for grid energy storage systems." IEEE Access 6 (2017): 13231-13260.
12. Su, Wencong, and Jianhui Wang. "Energy management systems in microgrid operations." The Electricity Journal 25, no. 8 (2012): 45-60.
13. Zhou, Bin, Wentao Li, Ka Wing Chan, Yijia Cao, Yonghong Kuang, Xi Liu, and Xiong Wang. "Smart home energy management systems: Concept, configurations, and scheduling strategies." Renewable and Sustainable Energy Reviews 61 (2016): 30-40.
14. Zhang, Huaguang, Yushuai Li, David Wenzhong Gao, and Jianguo Zhou. "Distributed optimal energy management for energy internet." IEEE Transactions on Industrial Informatics 13, no. 6 (2017): 3081-3097.
15. Doukas, Haris, Konstantinos D. Patlitzianas, Konstantinos Iatropoulos, and John Psarras. "Intelligent building energy management system using rule sets." Building and environment 42, no. 10 (2007): 3562-3569.
16. Bogdanovic, Dejan, Djordje Nikolic, and Ivana Ilic. "Mining method selection by integrated AHP and PROMETHEE method." *Anais da Academia Brasileira de Ciências* 84 (2012): 219-233.
17. Briggs, Th, P. L. Kunsch, and Bertrand Mareschal. "Nuclear waste management: an application of the multicriteria PROMETHEE methods." *European Journal of Operational Research* 44, no. 1 (1990): 1-10.
18. Abdullah, Lazim, Waimun Chan, and Alireza Afshari. "Application of PROMETHEE method for green supplier selection: a comparative result based on preference functions." *Journal of Industrial Engineering International* 15 (2019): 271-285.
19. Zhaoxu, Sun, and Han Min. "Multi-criteria decision making based on PROMETHEE method." In *2010 international conference on computing, control and industrial engineering*, vol. 1, pp. 416-418. IEEE, 2010.
20. Safari, Hossein, Maryam Sadat Fagheyi, Saadeh Sadat Ahangari, and M. Reza Fathi. "Applying PROMETHEE method based on entropy weight for supplier selection." *Business management and strategy* 3, no. 1 (2012): 97-106.
21. Animah, Isaac, and Mahmood Shafiee. "Maintenance strategy selection for critical shipboard machinery systems using a hybrid AHP-PROMETHEE and cost benefit analysis: a case study." *Journal of Marine Engineering & Technology* 20, no. 5 (2021): 312-323.
22. Deshmukh, S. C. "Preference ranking organization method of enrichment evaluation (promethee)." *International Journal of Engineering Science Invention* 2, no. 11 (2013): 28-34.
23. Polat, Gul. "Subcontractor selection using the integration of the AHP and PROMETHEE methods." *Journal of Civil Engineering and Management* 22, no. 8 (2016): 1042-1054.

24. Ozsahin, Dilber Uzun, Berna Uzun, Musa Sani Musa, Niyazi Şentürk, Fatih Veysel Nurçin, and Ilker Ozsahin. "Evaluating nuclear medicine imaging devices using fuzzy PROMETHEE method." *Procedia computer science* 120 (2017): 699-705.
25. PK Kanumarlapudi. (2024) Improving Data Governance with Advanced trade-off-Based Decision Models: A Comparative Analysis of Open Data Platform Implementations in Multiple Domains. SOJ Mater Sci Eng 10(2): 1-6. DOI: 10.15226/2473-3032/10/2/00183
26. Wang, Tien-Chin, Lisa Y. Chen, and Ying-Hsiu Chen. "Applying fuzzy PROMETHEE method for evaluating IS outsourcing suppliers." In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 3, pp. 361-365. IEEE, 2008.
27. PK Kanumarlapudi. "Improving Business Intelligence Reporting for Emerson Electric: A VIKOR-Based Comparative Evaluation of BI Tools" Journal of Business Intelligence and Data Analytics., 2025, vol. 3, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jbid.v3i2.267
28. Wu, Yunna, Meng Yang, Haobo Zhang, Kaifeng Chen, and Yang Wang. "Optimal site selection of electric vehicle charging stations based on a cloud model and the PROMETHEE method." *Energies* 9, no. 3 (2016): 157.
29. Bottero, Marta, Chiara D'Alpaos, and Alessandra Oppio. "Multicriteria evaluation of urban regeneration processes: an application of PROMETHEE method in Northern Italy." *Advances in Operations Research* 2018, no. 1 (2018): 9276075.
30. Verly, Céline, and Yves De Smet. "Some results about rank reversal instances in the PROMETHEE methods." *International Journal of Multicriteria Decision Making 71* 3, no. 4 (2013): 325-345.
31. PK Kanumarlapudi. "Big Data and Marketing Strategies in Banking: An SPSS Statistical Examination of Customer Engagement and Satisfaction"International Journal of Computer Science and Data Engineering., 2025, vol. 2, no. 3, pp. 1–6. doi: https://doi.org/10.55124/csdb.v2i3.252
32. Veza, Ivica, Stipo Celar, and Ivan Peronja. "Competences-based comparison and ranking of industrial enterprises using PROMETHEE method." *Procedia Engineering* 100 (2015): 445-449.
33. Kazan, Halim, Salih Özçelik, and Elif Haykir Hobikoğlu. "Election of deputy candidates for nomination with AHP-Promethee methods." *Procedia-Social and Behavioral Sciences* 195 (2015): 603-613.