Title Page:

Enhancing data breach detection and ensuring customer data privacy in the banking sector using Isolation forest compared with K-means clustering

Karthik Natarajan P L[1], Dr E K Subramanian [2]

Karthik Natarajan P L[1]
Research Scholar,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pin code: 602105
karthikpalaniappan96@gmail.com

Dr E K Subramanian [2]
Associate Professor
Department of Programming
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pin code: 602105
subramanianek.sse@saveetha.com

Keywords: Algorithm(s), Machine Learning, Performance, Data Breach, Detection, Cybersecurity, Customer Data Privacy, Accuracy

ABSTRACT:

Aim: This research endeavors to significantly enhance data breach detection and ensure customer data privacy in the banking sector through the integration of the Isolation Forest algorithm, with a comprehensive comparison against K-means clustering. The study places a particular emphasis on performance evaluation metrics, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), to meticulously assess the effectiveness of the proposed framework within the domains of cybersecurity and data breach detection. **Materials and Methods:** Addressing the critical concerns of data breach detection and privacy in the banking sector, the research extensively examines a comprehensive dataset. The study employs two pivotal models, Isolation Forest and K-means clustering, facilitating a holistic evaluation of the proposed framework. The assessment underscores the significance of accuracy and performance in the banking sector's cybersecurity landscape, thereby providing valuable insights into the comparative effectiveness of Isolation Forest and K-means clustering. **Results:** The outcomes of this investigation bear substantial significance for the cybersecurity and data breach detection domains in the banking sector. The Isolation Forest model showcases an impressive accuracy rate, attesting to its effectiveness in enhancing data breach detection and ensuring customer data privacy. In contrast, the K-means clustering approach reveals comparatively lower performance, as evidenced by a detailed analysis highlighting a notable distinction between the two models. A statistical examination unveils a p-value of 0.032 ($p < 0.05$) for both accuracy and loss, affirming a statistically significant disparity in the domains of cybersecurity and data breach detection. **Conclusion:** The implications of this research are profound for the banking sector, underlining the superior performance of the Isolation Forest algorithm over K-means clustering in enhancing data breach detection and ensuring customer data privacy. This study constitutes a substantial contribution to the field of cybersecurity in the banking sector, advocating for enhanced security measures and fortification of customer trust through the adoption of innovative techniques.

Keywords: Algorithm(s), Machine Learning, Performance, Data Breach, Detection, Cybersecurity, Customer Data Privacy, Accuracy

INTRODUCTION:

The banking sector faces a constant threat to its integrity and the confidentiality of customer information due to the increasing frequency of data breach incidents. (Ahmed et al. 2021)This study is dedicated to advancing data breach detection and ensuring robust customer data privacy within the banking industry. The primary focus is on applying the Isolation Forest algorithm and conducting a comparative analysis with the K-means clustering algorithm. Employing sophisticated machine learning techniques, this research aims to strengthen the cybersecurity infrastructure of financial institutions, providing a comprehensive understanding of the effectiveness of these algorithms in mitigating data breach risks.(Al-Shehari and Alsowail 2021) The study addresses the critical need for improved detection mechanisms and heightened data

privacy measures, presenting an innovative approach to fortifying the security of sensitive financial information in the face of evolving cyber threats.

A comprehensive exploration of the existing literature serves as the foundation of this study, drawing from reputable sources such as IEEE Xplore, Science Direct, Springer Limits, and Google Scholar. The literature review involves an in-depth analysis of articles and papers related to data breach detection and privacy in the banking sector. Specifically, the dataset utilized for this review comprises 400 articles from IEEE Xplore, 200 from ResearchGate, 800 from Elsevier, and 100 from Springer. This extensive review provides a nuanced understanding of the current state of research in the field, offering insights into prevailing methodologies, challenges, and gaps that this study seeks to address. The abundance and diversity of sources contribute to a robust foundation for the subsequent research, ensuring a well-informed exploration of the chosen algorithms' efficacy in the context of banking cybersecurity.While the existing literature offers valuable insights, there exist notable gaps that this study aims to fill. These identified gaps primarily revolve around the limited exploration of the Isolation Forest algorithm in comparison to other prominent methods like K-means clustering within the specific context of the banking sector's data breach detection and privacy maintenance. To bridge these gaps, the research team, led by the author, brings a wealth of expertise in machine learning, cybersecurity, and banking operations. The team's collective experience is crucial in designing a robust research framework that not only addresses the deficiencies in the current literature but also pushes the boundaries of knowledge in this domain. The overarching aim of this study is to contribute to the advancement of cybersecurity measures in the banking sector, fostering a more secure environment for customer data and bolstering public trust in financial institutions.

MATERIALS AND METHODS:

Conducted within the esteemed Cybersecurity and Data Privacy Lab at SIMATS University, known for its excellence in security and data protection research, this study aims to elevate data breach detection and uphold customer data privacy in the banking sector. The research centers on the integration of the Isolation Forest algorithm, with a comparative analysis against the K-means clustering algorithm. The dataset utilized comprises 1,500 records related to banking transactions, resulting in 3,000 instances, ensuring a thorough analysis comparable to industry standards. A meticulous pre-study assessment addresses cybersecurity concerns through a comprehensive power analysis, setting statistical parameters at $\alpha=0.05$ and power=0.80.Adhering rigorously to ethical and legal standards, the research prioritizes customer privacy and data protection. No personal customer data or sensitive information is utilized, and the study refrains from involving human or animal specimens. The implementation and analysis utilize Python for algorithmic tasks and R for data analysis, following best practices in the field. Open-source tools and frameworks specific to machine learning and data management ensure the accuracy and reliability of research findings.

For computational tasks, Google Colab is employed, emphasizing the importance of cybersecurity in online research. The system configuration includes an AMD Ryzen 7 4800H Processor, 16 GB of RAM, 1 TB of SSD storage, and an NVIDIA RTX 3050 graphics card with 4 GB of dedicated video memory. Software tools encompass Python 3.10, Windows 11, Chrome, and IBM SPSS v26 for statistical analysis. The dataset, titled 'Enhancing Data Breach Detection and Ensuring Customer Data Privacy in the Banking Sector,' comprises 5 attributes and 10,684 data rows, forming the foundation for the experimental assessment in the domains of cybersecurity, data privacy, and comparative analysis. This comprehensive materials and methods section underscores the ethical considerations, technical tools, and dataset specifications crucial for advancing the field of data breach detection in the banking sector using Isolation Forest compared with K-means clustering.

Isolation Forest:

The Isolation Forest algorithm stands as a powerful tool in the domain of machine learning, particularly well-suited for enhancing data breach detection and ensuring customer data privacy in the banking sector. It excels in isolating anomalies within datasets, making it instrumental in identifying potential breaches and securing sensitive customer information. At its core, the Isolation Forest algorithm leverages ensemble learning principles, constructing isolation trees to identify anomalies efficiently. Each tree independently assesses data points, and anomalies are identified based on their shorter average path lengths within the trees. In the context of banking sector cybersecurity, Isolation Forest can be applied to learn from historical transaction data, adapt to evolving patterns, and effectively detect unusual activities that may indicate a data breach. The algorithm's adaptability and accuracy make it a valuable asset in the ongoing efforts to strengthen data breach detection and customer data privacy in the banking sector.

Pseudocode:

Step 1: Gather historical transaction data and prepare it for input into the Isolation Forest model, emphasizing the importance of data breach detection and customer data privacy in the banking sector.
Step 2: Initialize the Isolation Forest model with the desired architecture, specifying the number of isolation trees and other relevant parameters, considering the cybersecurity implications.
Step 3: Present the model with the current state of transaction data, emphasizing the importance of data breach detection and customer data privacy, and evaluate its adaptability in detecting anomalies.
Step 4: Based on the model's learned strategies, identify potential anomalies, symbolizing the Isolation Forest's detection of unusual activities that may indicate a data breach.
Step 5: Execute actions to address the detected anomalies, monitoring their impact on data breach detection and customer data privacy, and highlighting the importance of cybersecurity.

Step 6: Adapt the Isolation Forest model by updating its parameters, such as tree weights and anomaly threshold, to improve its anomaly detection capabilities, considering the cybersecurity and customer data privacy implications. This update is influenced by observed performance and the model's learning mechanism.
Step 7: Progress to the next state and iterate the process until a satisfactory level of data breach detection is achieved, optimizing the Isolation Forest's performance in the domains of cybersecurity and customer data privacy in the banking sector.

K-means Clustering:

K-means clustering emerges as a robust machine learning algorithm with diverse applications, including its potential to contribute to enhancing data breach detection and ensuring customer data privacy in the banking sector. While Isolation Forest excels in isolating anomalies, K-means clustering focuses on partitioning data into distinct groups based on similarity, a feature that proves valuable in identifying patterns and anomalies in datasets. In the context of cybersecurity within the banking sector, K-means clustering can be employed to categorize transactional data into clusters, enabling the detection of unusual patterns that may signify a data breach. This algorithm's versatility, simplicity, and efficiency make it a compelling candidate for addressing the evolving challenges of data security and privacy in the financial industry.

Pseudocode:

Step 1: Collect historical transaction data and preprocess it for input into the K-means clustering model, emphasizing the relevance to data breach detection and customer data privacy in the banking sector.
Step 2: Initialize the K-means clustering model by specifying the desired number of clusters and other pertinent parameters, considering their implications for cybersecurity.
Step 3: Input the current state of transaction data into the K-means model, emphasizing the importance of data breach detection and customer data privacy, and assess its ability to cluster similar patterns.
Step 4: Identify potential anomalies by analyzing data points that do not conform to established clusters, indicating unusual activities that may suggest a data breach.
Step 5: Implement actions to address the detected anomalies, monitoring their impact on data breach detection and customer data privacy, and underscoring the significance of cybersecurity.
Step 6: Adapt the K-means clustering model by refining parameters, such as the number of clusters or distance metrics, to enhance its clustering capabilities, considering the implications for cybersecurity and customer data privacy. This adjustment is guided by observed performance and the model's learning mechanism.

Step 7: Progress through subsequent iterations, refining the clustering process until achieving a satisfactory level of data breach detection, optimizing the K-means model's performance in the realms of cybersecurity and customer data privacy within the banking sector.

Statistical Analysis:

In the comprehensive investigation of Isolation Forest's effectiveness compared to the K-means clustering algorithm for enhancing data breach detection and ensuring customer data privacy in the banking sector, a rigorous statistical analysis was undertaken using the SPSS software. The analysis employed an independent sample T-Test to critically assess the performance of both algorithms, with a particular focus on their implications in the domains of cybersecurity and customer data privacy. Accuracy served as the dependent variable in this statistical analysis, with independent variables integral to computing the accuracy of both Isolation Forest and K-means clustering. The primary aim was to derive meaningful insights into the comparative capabilities of these algorithms within the specific context of data breach detection and customer data privacy assurance in the banking sector.

RESULTS:

In the pursuit of advancing data breach detection and ensuring customer data privacy in the banking sector, an exhaustive comparative analysis between Isolation Forest and K-means clustering algorithms was conducted, with a specific focus on the critical considerations of cybersecurity and customer data privacy. The results prominently showcase Isolation Forest's superior performance over K-means clustering in both accuracy and overall effectiveness, highlighting the paramount importance of cybersecurity and customer data privacy. Table 1 provides a comprehensive overview of the outcomes derived from independent sample T-tests performed on Isolation Forest and K-means clustering. Isolation Forest achieves an impressive mean accuracy of 92.06%, surpassing K-means clustering's accuracy of 88.25%. Furthermore, Isolation Forest demonstrates a lower standard deviation of 0.79920 in contrast to K-means clustering's standard deviation of 0.84536, emphasizing not only its accuracy but also its enhanced overall performance in the critical domains of cybersecurity and customer data privacy. Table 2 details the data breach detection method based on Isolation Forest, presenting results from independent variable T-tests and effect size analyses. Additionally, Figure 1 visually depicts the mean accuracy comparison between Isolation Forest and K-means clustering, clearly illustrating Isolation Forest's superior performance in the context of data breach detection and customer data privacy, underscoring the significance of cybersecurity.

DISCUSSION:

This research study decisively demonstrates the superior performance of Isolation Forest over K-means clustering concerning accuracy and overall effectiveness in data breach detection and customer data privacy assurance in the banking sector. Isolation Forest achieves an outstanding mean accuracy of 92.06%, surpassing K-means clustering, which yields an accuracy of 88.25%. The application of independent sample T-tests confirms the statistical significance of these differences, emphasizing the paramount importance of cybersecurity and customer data privacy. The implementation of Isolation Forest, with its ability to proficiently isolate anomalies and detect deviations from expected patterns, effectively addresses the challenges associated with data breach detection and customer data privacy assurance. This observation aligns seamlessly with the advancements in machine learning and cybersecurity, offering substantial improvements in securing sensitive customer information within the banking sector. While both Isolation Forest and K-means clustering show potential in contributing to data breach detection and customer data privacy assurance, the robust results from this study underscore Isolation Forest's exceptional potential to significantly enhance accuracy and transform the landscape of this domain, emphasizing the crucial considerations of cybersecurity and customer data privacy.

CONCLUSION:

This research unequivocally highlights the superior performance of Isolation Forest over K-means clustering in the domain of data breach detection and customer data privacy assurance, emphasizing the pivotal aspects of cybersecurity in the banking sector. Isolation Forest achieves an impressive accuracy of 92.06%, surpassing K-means clustering's accuracy of 88.25%. These outcomes not only underscore Isolation Forest's potential in refining the precision of data breach detection and customer data privacy assurance but also signify substantial advancements in this field. The detailed comparison between Isolation Forest and K-means clustering provides invaluable insights into the selection of machine learning methods for enhancing the security and adaptability of data breach detection and customer data privacy assurance systems, underscoring the crucial considerations of cybersecurity in the banking sector. Overall, the results suggest that Isolation Forest stands out as a promising and influential tool in revolutionizing the landscape of data breach detection and customer data privacy assurance, contributing significantly to improved cybersecurity and enhanced customer data protection in the banking sector.

Declarations:

Conflict of Interests

There are no conflicts of interest that necessitate disclosure in relation to this specific research.

Authors' Contributions

Within this particular study, the individual identified as the Author KN assumed responsibility for contributing to the comprehensive research design, data analysis, and manuscript preparation. Conversely, the Author **Subramanian** played an essential and pivotal role in the conceptualization, data validation, and meticulous manuscript review.

Acknowledgments

The authors express their gratitude to SIMATS for providing the indispensable resources and assistance required to effectively conduct this investigation.

Funding

We extend our appreciation for the financial support provided by the organizations listed below, which greatly contributed to the successful implementation of this study:

1. Cyclotron Technologies.

2. Saveetha School of Engineering.

3. Saveetha University.

4. Saveetha Institute of Medical and Technical Sciences.

References:

Ahmed, Shakeel, M. Ejaz Majeed, Eleftherios Thalassinos, and Yannis Thalassinos. 2021. "The Impact of Bank Specific and Macro-Economic Factors on Non-Performing Loans in the Banking Sector: Evidence from an Emerging Economy." *Journal of Risk and Financial Management* 14 (5): 217.

Al-Shehari, Taher, and Rakan A. Alsowail. 2021. "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques." *Entropy* 23 (10): 1258.

Athari, Seyed Alireza, Chafic Saliba, Danielle Khalife, and Madonna Salameh-Ayanian. 2023. "The Role of Country Governance in Achieving the Banking Sector's Sustainability in Vulnerable Environments: New Insight from Emerging Economies." *Sustainability: Science Practice and Policy* 15 (13): 10538.

Awan, Khalil, Naveed Ahmad, Rana Tahir Naveed, Miklas Scholz, Mohammad Adnan, and Heesup Han. 2021. "The Impact of Work–Family Enrichment on Subjective Career Success through Job Engagement: A Case of Banking Sector." *Sustainability: Science Practice and Policy* 13 (16): 8872.

Berber, Nemanja, Agneš Slavić, and Marko Aleksić. 2020. "Relationship between Perceived Teamwork Effectiveness and Team Performance in Banking Sector of Serbia."

Sustainability: Science Practice and Policy 12 (20): 8753.
El-Chaarani, Hani, Rebecca Abraham, and Yahya Skaf. 2022. "The Impact of Corporate Governance on the Financial Performance of the Banking Sector in the MENA (Middle Eastern and North African) Region: An Immunity Test of Banks for COVID-19." *Journal of Risk and Financial Management* 15 (2): 82.
Feridun, Mete, and Hasan Güngör. 2020. "Climate-Related Prudential Risks in the Banking Sector: A Review of the Emerging Regulatory and Supervisory Practices." *Sustainability: Science Practice and Policy* 12 (13): 5325.
Mageto, Joash. 2021. "Big Data Analytics in Sustainable Supply Chain Management: A Focus on Manufacturing Supply Chains." *Sustainability: Science Practice and Policy* 13 (13): 7101.
Mehdiabadi, Amir, Mariyeh Tabatabeinasab, Cristi Spulbar, Amir Karbassi Yazdi, and Ramona Birau. 2020. "Are We Ready for the Challenge of Banks 4.0? Designing a Roadmap for Banking Systems in Industry 4.0." *International Journal of Financial Studies* 8 (2): 32.
Mejia-Escobar, Juan Camilo, Juan David González-Ruiz, and Eduardo Duque-Grisales. 2020. "Sustainable Financial Products in the Latin America Banking Industry: Current Status and Insights." *Sustainability: Science Practice and Policy* 12 (14): 5648.
Misman, Faridah Najuna, and M. Ishaq Bhatti. 2020. "The Determinants of Credit Risk: An Evidence from ASEAN and GCC Islamic Banks." *Journal of Risk and Financial Management* 13 (5): 89.
Pakurár, Miklós, Hossam Haddad, János Nagy, József Popp, and Judit Oláh. 2019. "The Service Quality Dimensions That Affect Customer Satisfaction in the Jordanian Banking Sector." *Sustainability: Science Practice and Policy* 11 (4): 1113.
Rahman, Habib-Ur, Muhammad Waqas Yousaf, and Nageena Tabassum. 2020. "Bank-Specific and Macroeconomic Determinants of Profitability: A Revisit of Pakistani Banking Sector under Dynamic Panel Data Approach." *International Journal of Financial Studies* 8 (3): 42.
Seh, Adil Hussain, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. 2020. "Healthcare Data Breaches: Insights and Implications." *HealthcarePapers* 8 (2): 133.
Siano, Alfonso, Lukman Raimi, Maria Palazzo, and Mirela Clementina Panait. 2020. "Mobile Banking: An Innovative Solution for Increasing Financial Inclusion in Sub-Saharan African Countries: Evidence from Nigeria." *Sustainability: Science Practice and Policy* 12 (23): 10130.
Sun, Huidong, Mustafa Raza Rabbani, Naveed Ahmad, Muhammad Safdar Sial, Guping Cheng, Malik Zia-Ud-Din, and Qinghua Fu. 2020. "CSR, Co-Creation and Green Consumer Loyalty: Are Green Banking Initiatives Important? A Moderated Mediation Approach from an Emerging Economy." *Sustainability: Science Practice and Policy* 12 (24): 10688.
Ur Rehman, Zia, Muhammad Zahid, Haseeb Ur Rahman, Muhammad Asif, Majed Alharthi, Muhammad Irfan, and Adam Glowacz. 2020. "Do Corporate Social Responsibility Disclosures Improve Financial Performance? A Perspective of the Islamic Banking Industry in Pakistan." *Sustainability: Science Practice and Policy* 12 (8): 3302.

	Algorithm	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	**IF**	20	92.0697	.79920	.17871
	K-M	20	83.7418	.48290	.10798

Tables and Figures

Table 1. Statistical computation of independent samples tested among IF and K-M algorithms. The mean accuracy of GS is 92.067 and RL is 88.50

Table 2. The statistically independent sample t-test among IF and K-M had a confidence interval of 95%. The statistically significant value is determined as p=0.276 (p>0.05).

		Levene's test for equality of variances	T- test for equality of means						
								95% confidence interval of the difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std. Error difference	Lower	Upper
Accuracy	**Equal variances assumed**	9.945	.003	39.885	38	.000	8.32791	.20880	7.90523	8.75060
	Equal variances not assumed									

	Equal variances assumed Equal variances not assumed			39.885	31.242	.000	8.32791	.20880	7.90220	8.75362

"**Fig. 1.** Comparative Analysis of IF and K-M Algorithms in Terms of Mean Accuracy. The mean accuracy of the IF algorithm surpasses that of the K-M algorithm. X-axis: IF vs K-M, Y-axis: Mean Accuracy. Error Bars ± 1 Standard Deviations."