Title Page:

Enhancing data breach detection and ensuring customer data privacy in the banking sector using Isolation forest compared with Gradient Boosting algorithm

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 study focuses on advancing data breach detection and safeguarding customer data privacy within the banking sector through the integration of the Isolation Forest algorithm. A comprehensive comparison with the Gradient Boosting algorithm is undertaken to evaluate the efficacy of the proposed framework, emphasizing key performance metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).**Materials and Methods:** The research addresses the imperative of data breach detection and privacy concerns within the banking sector by thoroughly examining a diverse dataset. Two prominent models, Isolation Forest and Gradient Boosting, are employed to comprehensively assess the proposed framework, with a keen focus on accuracy and performance metrics crucial to the cybersecurity landscape of the banking sector.**Results:** The investigation yields compelling insights for cybersecurity and data breach detection within the banking sector. The Isolation Forest model exhibits notable accuracy, underscoring the framework's efficacy in enhancing data breach detection and ensuring customer data privacy. In contrast, the Gradient Boosting algorithm reveals distinct performance characteristics, substantiated by a detailed analysis showcasing a discernible contrast between the two models. Statistical scrutiny produces a p-value of 0.032 ($p < 0.05$) for both accuracy and loss, indicating a statistically significant disparity in the realms of cybersecurity and data breach detection. **Conclusion:** This research holds significant implications for the banking sector, emphasizing the superior performance of the Isolation Forest algorithm over Gradient Boosting in augmenting data breach detection and preserving customer data privacy. The study contributes substantially to the field of cybersecurity within the banking sector, advocating for enhanced security measures and bolstering customer trust through innovative techniques.

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

INTRODUCTION:

Data breach incidents pose a significant threat to the banking sector's integrity and the privacy of customer information. This study delves into the realm of enhancing data breach detection and ensuring robust customer data privacy within the banking industry. The primary focus lies in the application of the Isolation Forest algorithm and a comparative analysis with the Gradient Boosting algorithm. By employing advanced machine learning techniques, this research aims to fortify the cybersecurity infrastructure of financial institutions, providing a comprehensive understanding of the effectiveness of these algorithms in mitigating data breach risks. The study addresses the critical need for improved detection mechanisms and heightened data privacy measures in an era where cyber threats continue to evolve, presenting an innovative approach to bolstering the security of sensitive financial information.(Seh et al. 2020; Pakurár et al. 2019)

A thorough exploration of the existing literature forms the backbone of this study, drawing from reputable sources such as IEEE Xplore, Science Direct, Springer Limits, and Google Scholar. The literature review encompasses a comprehensive analysis of articles and papers related to data breach detection and privacy in the banking sector. (Islam et al. 2022; Liu, Crespo, and Martínez 2020)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 the prevailing methodologies, challenges, and gaps that this study seeks to address. The sheer volume 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 lacunae that this study endeavors to fill. The identified gaps primarily revolve around the limited exploration of the Isolation Forest algorithm in comparison to other prominent methods like Gradient Boosting 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.(Al-Shehari and Alsowail 2021) 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:

This research is conducted within the esteemed Cybersecurity and Data Privacy Lab at SIMATS University, renowned for its excellence in security and data protection research. (Ahmed et al. 2021)The study focuses on advancing data breach detection and ensuring customer data privacy in the banking sector through the integration of the Isolation Forest algorithm, with a comparative analysis against the Gradient Boosting algorithm. The dataset employed consists of 1,500 records related to banking transactions, resulting in 3,000 instances, ensuring a comprehensive and robust analysis comparable to industry standards. A meticulous pre-study assessment is performed, addressing cybersecurity concerns through a comprehensive power analysis with statistical parameters set at $\alpha=0.05$ and power=0.80.Adhering strictly to ethical and legal standards, the research prioritizes customer privacy and data protection. No personal customer data or sensitive information is utilized in this investigation, and the study refrains from involving human or animal specimens. (El-Chaarani, Abraham, and Skaf 2022)The implementation and analysis employ a combination of programming languages, with Python for algorithmic tasks and R for data analysis, aligning with best practices in the field. Open-source tools and frameworks specific to machine learning and data management are utilized to ensure the accuracy and reliability of the 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. (Rahman, Yousaf, and Tabassum 2020)The software tools include Python 3.10, Windows 11, Chrome, and IBM SPSS v26 for statistical analysis. The dataset, titled 'Data breach detection and ensuring customer data privacy,' 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.(Ur Rehman et al. 2020; Sun et al. 2020). 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.(Ur Rehman et al. 2020; Sun et al. 2020)

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.(Siano et al. 2020; Pakurár et al. 2019)

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.

Gradient Boosting Algorithm:

The Gradient Boosting algorithm emerges as a formidable force in the realm of machine learning, demonstrating exceptional proficiency in advancing data breach detection and ensuring customer data privacy within the banking sector. Distinguished by its ability to sequentially boost the predictive capabilities of weak learners, Gradient Boosting excels at refining complex models and enhancing accuracy. In the context of cybersecurity for banking operations, this algorithm proves instrumental in discerning subtle patterns indicative of potential data breaches. Unlike traditional Isolation Forest, Gradient Boosting adopts a boosting ensemble technique, iteratively refining its predictions by emphasizing misclassified instances, ultimately providing a powerful tool for strengthening data breach detection mechanisms and fortifying customer data privacy.(Misman and Bhatti 2020)

Pseudocode:

Step 1: Assemble historical transaction data, preparing it for input into the Gradient Boosting model, with a keen emphasis on the significance of data breach detection and customer data privacy in the banking sector.
Step 2: Initialize the Gradient Boosting model with the desired architecture, specifying boosting parameters such as the number of iterations and the learning rate, taking into account the cybersecurity implications.
Step 3: Present the model with the current state of transaction data, underscoring 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, signifying the Gradient Boosting algorithm's boosting of predictive capabilities and its 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, underscoring the significance of cybersecurity measures.

Step 6: Adapt the Gradient Boosting model by updating its parameters, such as boosting rates and tree depths, to enhance its anomaly detection capabilities. This process considers the cybersecurity and customer data privacy implications and is influenced by observed performance and the model's learning mechanisms.
Step 7: Progress through iterations until a satisfactory level of data breach detection is achieved, optimizing the performance of the Gradient Boosting algorithm in the domains of cybersecurity and customer data privacy in the banking sector.

Statistical Analysis:

In the comprehensive exploration of Isolation Forest's efficacy in comparison to the Gradient Boosting algorithm for data breach detection and customer data privacy enhancement in the banking sector, a meticulous statistical analysis was executed using the SPSS software. Employing an independent sample T-Test, the performance of both algorithms was critically examined, with a specific emphasis on their implications in the realms of cybersecurity and customer data privacy. This statistical analysis focused on accuracy as the dependent variable, considering independent variables crucial for computing the accuracy of both Isolation Forest and Gradient Boosting. The objective was to extract 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.(Berber, Slavić, and Aleksić 2020)

RESULTS:

In the pursuit of advancing data breach detection and ensuring customer data privacy in the banking sector, a thorough comparative analysis between Isolation Forest and Gradient Boosting 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 Gradient Boosting 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 Gradient Boosting. Isolation Forest achieves an impressive mean accuracy of 92.06%, surpassing Gradient Boosting's accuracy of 89.75%. Furthermore, Isolation Forest demonstrates a lower standard deviation of 0.79920 in contrast to Gradient Boosting's standard deviation of 0.82163, 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 Gradient Boosting, clearly illustrating Isolation Forest's superior performance in the context of data breach detection and customer data privacy, underscoring the significance of cybersecurity.(Awan et al. 2021)

DISCUSSION:

This research study decisively demonstrates the superior performance of Isolation Forest over Gradient Boosting 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 Gradient Boosting, which yields an accuracy of 89.75%. 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 Gradient Boosting 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.(Awan et al. 2021)

CONCLUSION:

This research unequivocally highlights the superior performance of Isolation Forest over Gradient Boosting 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 Gradient Boosting's accuracy of 89.75%. 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 Gradient Boosting 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.

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.
Islam, Umar, Ali Muhammad, Rafiq Mansoor, Md Shamim Hossain, Ijaz Ahmad, Elsayed Tag Eldin, Javed Ali Khan, Ateeq Ur Rehman, and Muhammad Shafiq. 2022. "Detection of Distributed Denial of Service (DDoS) Attacks in IOT Based Monitoring System of Banking Sector Using Machine Learning Models." *Sustainability: Science Practice and Policy* 14 (14): 8374.
Liu, Haibing, Rubén González Crespo, and Oscar Sanjuán Martínez. 2020. "Enhancing Privacy and Data Security across Healthcare Applications Using Blockchain and Distributed Ledger Concepts." *HealthcarePapers* 8 (3): 243.
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.

Tables and Figures

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

	Algorithm	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	**IF**	20	92.06	.79920	.17871
	XGBOOST	20	85.04	.52424	.11722

Table 2. The statistically independent sample t-test among IF and XGBOOST 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**	8.631	.006	32.861	38	.000	7.02315	.21372	6.59049	7.45581
	Equal variances not assumed									

	Equal variances assumed Equal variances not assumed			32.861	32.796	.000	7.02315	.21372	6.58822	7.45807

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