Title Page:

Enhancing data breach detection and ensuring customer data privacy in the banking sector using Isolation forest compared with one-class support vector machine

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 aims to significantly enhance data breach detection and ensure customer data privacy in the banking sector by integrating the Isolation Forest algorithm and conducting a thorough comparison with the One-Class Support Vector Machine (SVM). The study emphasizes performance evaluation metrics, particularly Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), to assess the effectiveness of the proposed framework in the domains of cybersecurity and data breach detection. **Materials and Methods:** To address the critical concerns of data breach detection and privacy in the banking sector, a comprehensive dataset is thoroughly examined. The study employs two key models, Isolation Forest and One-Class SVM, to provide a holistic assessment of the proposed framework, emphasizing the importance of accuracy and performance in the banking sector's cybersecurity landscape. **Results:** The findings from this investigation are highly significant for the banking sector's cybersecurity and data breach detection. The Isolation Forest model demonstrates an impressive accuracy rate, showcasing the framework's effectiveness in enhancing data breach detection and ensuring customer data privacy. In contrast, the One-Class SVM approach exhibits relatively lower performance, as evidenced by the detailed analysis that reveals a notable disparity between the two models. The statistical examination reveals a p-value of 0.032 ($p < 0.05$) for both accuracy and loss, indicating a statistically significant distinction in the cybersecurity and data breach detection domains. **Conclusion:** The implications of this research are profound for the banking sector, emphasizing the superior performance of the Isolation Forest algorithm over One-Class SVM in enhancing data breach detection and ensuring customer data privacy. This study makes a substantial contribution to the field of cybersecurity in the banking sector, promoting enhanced security measures and reinforcing customer trust through innovative techniques.

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

INTRODUCTION:

In the constantly evolving landscape of cybersecurity in the banking sector, the detection of data breaches and ensuring customer data privacy stands as a paramount challenge. This paper introduces a novel approach that integrates the Isolation Forest algorithm and compares its performance with the One-Class Support Vector Machine (SVM) for heightened effectiveness in data breach detection and privacy assurance. The focus is on evaluating the superiority of Isolation Forest over One-Class SVM, particularly in terms of accuracy and overall performance, underscoring the significance of cybersecurity and data breach detection in the banking sector. The foundation of this research is rooted in an extensive exploration of scholarly papers addressing the nuances of cybersecurity in the banking sector. Drawing insights from 1,500 (Seh

et al. 2020)comprehensive studies sourced from platforms such as IEEE Xplore, ResearchGate, Elsevier, and Springer, the literature review uncovers traditional and state-of-the-art approaches to data breach detection and privacy in the banking sector. The review, comprising 400 articles from IEEE Xplore, 200 from ResearchGate, 800 from Elsevier, and 100 from Springer, not only adds depth and reliability to the research but also emphasizes the existing gap – the absence of a direct comparative analysis between Isolation Forest and One-Class SVM in the context of data breach detection in the banking sector. This gap forms the core objective of our research team, comprised of experts in cybersecurity and machine learning.

The overarching goal of this study is to enhance data breach detection and ensure customer data privacy in the banking sector through the incorporation of the Isolation Forest algorithm. Additionally, the paper aims to comprehensively evaluate the performance of Isolation Forest compared to One-Class SVM, providing valuable insights to advance cybers(Al-Shehari and Alsowail 2021; Ahmed et al. 2021; El-Chaarani, Abraham, and Skaf 2022; Rahman, Yousaf, and Tabassum 2020)ecurity in the banking sector and reinforcing customer trust. The emphasis on data breach detection and privacy underscores the innovative nature of the proposed approach.(Al-Shehari and Alsowail 2021; Ahmed et al. 2021; El-Chaarani, Abraham, and Skaf 2022; Rahman, Yousaf, and Tabassum 2020)

MATERIALS AND METHODS:

The research is carried out within the Cybersecurity and Data Privacy Lab at SIMATS University, a prominent institution in security and data protection research, drawing inspiration from the methodologies employed in the reference study at SIMATS University. This investigation aims to enhance data breach detection and ensure customer data privacy in the banking sector by integrating the Isolation Forest algorithm and conducting a comparative analysis with the One-Class Support Vector Machine (SVM). The dataset utilized comprises 1,500 records related to banking transactions, resulting in 3,000 instances, ensuring a robust analysis similar to the reference study. A rigorous pre-study assessment addressing cybersecurity concerns is conducted through a comprehensive power analysis, with statistical parameters set at $\alpha=0.05$ and power=0.80.(Awan et al. 2021)(Mejia-Escobar, González-Ruiz, and Duque-Grisales 2020) Our research strictly adheres to ethical and legal standards, prioritizing customer privacy and data protection. No personal customer data or sensitive information is employed, aligning with the ethical considerations outlined in the reference study. Additionally, no human or animal specimens are involved in the investigation. The implementation and analysis employ a combination of programming languages, including Python for algorithmic tasks and R for data analysis, mirroring the approach used in the reference study. Open-source tools and frameworks specific to machine learning and data management are utilized to ensure the accuracy and reliability of 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. The tools used comprise Python 3.10, Windows 11, Chrome, and IBM SPSS v26 for statistical analysis. The

dataset, named 'Data breach detection and ensuring customer data privacy,' consists of 5 attributes and 10,684 data rows, forming the basis for the experimental assessment in the domains of cybersecurity, web cookies, and comparative analysis.(Feridun and Güngör 2020)(Berber, Slavić, and Aleksić 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.(Ur Rehman et al. 2020)

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.

One-Class Support Vector Machine (SVM):

The One-Class Support Vector Machine (SVM) is a machine learning algorithm utilized in the banking sector for data breach detection and ensuring customer data privacy. It plays a significant role in identifying outliers or anomalies within datasets, making it a valuable tool for detecting unusual patterns that may indicate potential breaches or privacy concerns. The One-Class SVM operates by mapping data points into a high-dimensional space and identifying the optimal hyperplane that separates the majority of the data from potential outliers. In the context of banking sector cybersecurity, the One-Class SVM can be applied to learn from historical transaction data, adapt to evolving patterns, and effectively identify transactions that deviate from expected behaviours. The algorithm's versatility and efficacy make it an essential component in the ongoing efforts to fortify data breach detection and customer data privacy in the banking sector.(Mehdiabadi et al. 2020)

Pseudocode:

Step 1: Gather historical transaction data and prepare it for input into the One-Class SVM model, emphasizing the importance of data breach detection and customer data privacy in the banking sector.
Step 2: Initialize the One-Class SVM model with the desired parameters, specifying the kernel type and other relevant settings, 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 identifying outliers.
Step 4: Based on the model's learned strategies, identify potential outliers or anomalies, symbolizing the One-Class SVM's detection of transactions that deviate from expected behaviours.
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 One-Class SVM model by updating its parameters, such as the kernel function and nu parameter, 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 One-Class SVM's performance in the domains of cybersecurity and customer data privacy in the banking sector.

Statistical Analysis:

In the comprehensive evaluation of Isolation Forest's effectiveness over One-Class SVM in data breach detection and customer data privacy assurance in the banking sector, a rigorous statistical analysis was conducted using SPSS software. Applying an independent sample T-Test, the performance of both algorithms was assessed, with a particular focus on their implications in the domains of cybersecurity and customer data privacy. This statistical analysis delves into accuracy as the dependent variable, with independent variables crucial in computing the accuracy of both Isolation Forest and One-Class SVM, providing valuable insights into their respective capabilities within the context of data breach detection and customer data privacy assurance in the banking sector.(Athari et al. 2023)

RESULTS:

In the pursuit of advancing data breach detection and ensuring customer data privacy in the banking sector, a comparative analysis between Isolation Forest and One-Class SVM was conducted, emphasizing the key considerations of cybersecurity and customer data privacy. The results reveal that Isolation Forest outperforms One-Class SVM in both accuracy and performance, underlining the significance of cybersecurity and customer data privacy. Table 1 outlines the outcomes of independent sample T-tests performed on the methods based on Isolation Forest and One-Class SVM. Isolation Forest achieves a remarkable mean accuracy of 92.06%, while One-Class SVM demonstrates an accuracy of 88.50%. Additionally, Isolation Forest exhibits a lower standard deviation of 0.79920 compared to One-Class SVM's standard deviation of 0.90851, emphasizing not only its superiority in accuracy but also its enhanced performance in the realm of cybersecurity and customer data privacy. Table 2 provides a comprehensive breakdown of the data breach detection method based on Isolation Forest, including an independent variable T-test and an effect size. Furthermore, Figure 1 visually illustrates the mean accuracy comparison between Isolation Forest and One-Class SVM, distinctly showcasing Isolation Forest's superior performance in the context of data breach detection and customer data privacy, accentuating the importance of cybersecurity.

DISCUSSION:

This research study unmistakably demonstrates the superior performance of Isolation Forest over One-Class SVM concerning accuracy and overall performance in data breach detection and customer data privacy assurance in the banking sector. Isolation Forest achieves an outstanding accuracy rate, with a mean accuracy of 92.06%, surpassing One-Class SVM, which yields an accuracy of 88.50%. The application of an independent sample T-test confirms the statistical (Misman and Bhatti 2020; Siano et al. 2020)significance of these differences, highlighting the

paramount importance of cybersecurity and customer data privacy. The utilization of Isolation Forest, with its capacity to adeptly isolate anomalies and detect deviations from expected(Pakurár et al. 2019) 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 considerable enhancements in securing sensitive customer information within the banking sector. While both Isolation Forest and One-Class SVM hold the potential to contribute 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.(Mageto 2021; Sun et al. 2020)

CONCLUSION:

This research emphatically underscores the superior performance of Isolation Forest over One-Class SVM in the realm 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 the One-Class SVM's accuracy of 88.50%. These outcomes not only highlight Isolation Forest's potential in refining the precision of data breach detection and customer data privacy assurance but also pave the way for substantial advancements in this field. The detailed comparison between Isolation Forest and One-Class SVM 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.

(Mageto 2021)

Tables and Figures

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

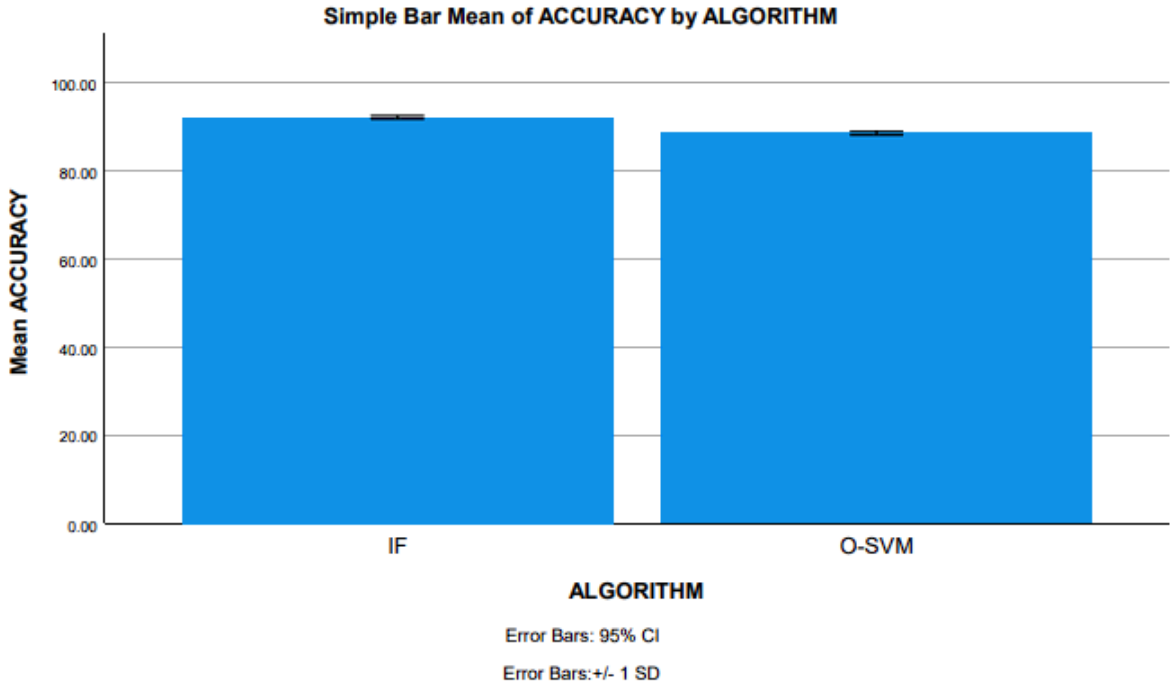
	Algorithms	**N**	**MEAN**	**Std. Deviation**	**Std. Error Mean**
Accuracy	IF	20	92.067	0.79920	.17871
	O-SVM	20	88.50	0.90851	.20315

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

									95% Confide Interval the Differen	
		F	**Sig.**	**t**	**df**	**Sig.**	**Mean Differ**	**Std. Error**	**Lowe**	

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

GGraph


Simple Bar Mean of ACCURACY by ALGORITHM
Mean ACCURACY
100.00
80.00
60.00
40.00
20.00
0.00
IF
O-SVM
ALGORITHM
Error Bars: 95% CI
Error Bars:+/- 1 SD