MULTIDISCIPLINARY SCIENTIFIC RESEARCH
BJMSR VOL 11 NO 2 (2026) P-ISSN 2687-850X E-ISSN 2687-8518
Available online at https://www.cribfb.com
Journal homepage: https://www.cribfb.com/journal/index.php/BJMSR
Published by CRIBFB, USA

QUANTIFYING THE VULNERABILITY OF CREDIT CARD FRAUD DETECTION SYSTEMS TO GRADIENT BASED ADVERSARIAL EVASION

Nabil El Kadhi [(a)1] *Marwan Alshar'e* [(b)] *Rifat O. Shannak* [(c)] *Nour Eldin Elshaiekh Osman* [(d)] *Basel Bani-Ismail* [(e)]

[(a)]*Associate Professor, VPAA and Computer Science Department, Applied Science University, Manama, Bahrain; E-mail: Nabil.elkadhi@asu.edu.bh*
[(b)]*Associate Professor, Faculty of Computing and IT, Sohar University, Sohar, Sultanate of Oman; E-mail: mshare@su.edu.om*
[(c)]*Professor, Business School, Jordan University, Amman, Jordan, E-mail: rshannak@ju.edu.jo*
[(d)]*Associate Professor, Sultan Qaboos University, College of Arts and Social Sciences, Department of Information Studies, Sultanate of Oman; E-mail: n.osman@squ.edu.om*
[(e)]*Assistant Professor, Department of Software Engineering, Faculty of Information Technology, Applied Science Private University, Amman, Jordan ;E-mail: b_ismail@asu.edu.jo*

ARTICLE INFO

Article History:

Received: 17th October 2025
Reviewed & Revised: 17th October 2025 to 17th April 2026
Accepted: 25th April 2026
Published: 29th April 2026

Keywords:

AI Security, Adversarial ML, Financial Cyber Resilience, Stealth Fraud Patterns, Feature Perturbation Analysis, Fast Gradient Method (FGM), Electronic Payment Security

JEL Classification Codes:

G21, K42, O33, C45

Peer-Review Model:

External peer review was done through double-blind method.

ABSTRACT

In financial cybersecurity, fraud detection is at the forefront of the battle against cybercriminals and deep learning models are the state-of-the-art method for real-time detection; yet their reliance on fixed feature distributions renders them inherently susceptible to adversarial attacks. With the rise of generative artificial intelligence (AI) in obfuscating fraudulent transactions, conventional performance indicators (accuracy and F1-score) do not reflect the cybersecurity resilience of these models. This research explores the Credit Card Fraud Detection dataset, comprising more than 550,000 anonymized records. The study uses a functional deep learning model as a benchmark and introduces the Fast Gradient Method (FGM) to simulate white box attacks on the latent features . The experiment results demonstrate a high detection accuracy of 99.85% under normal operating conditions, which dramatically reduces to 92.75% when card transactions are slightly modified by adversarial disturbances at an epsilon value of 0.1. Numerical evidence reveals a clear 7.10% "Security Gap," depicting a substantial degree of vulnerability, where transactions are mathematically reclassified from fraudulent to legitimate. Notably, our research suggests a non-linear "safety cliff" in the decrease of detection accuracy, where the model's reliability completely fails as the adversarial strength reaches 0.15 or higher. In addition, Principal Component Analysis demonstrates that gradient attacks have successfully distorted the latent features of fraudulent transactions, resulting in the movement of highly suspicious transactions towards the concentrated area of legitimate transactions, hence evading automated security screening while preserving the overall statistical distribution of the data.

INTRODUCTION

The financial sector is in the midst of a dramatic digital shift, in which real-time transaction processing is expected for consumer convenience. But with this convenience comes a greater opportunity for cyber-criminals to execute nefarious schemes, where credit card fraud detection is a linchpin of financial stability (Dupont et al., 2024). Whereas rules-based systems were once adequate, the 2022-2026 era has witnessed the rise of "AI-driven fraud," where hackers use generative models to simulate legitimate consumer behaviour patterns with remarkable accuracy (Chelliah et al., 2025). The impact of this research is the pressing need to strike a balance between classification accuracy and cyber-resilience. The scientific challenge relates to the vulnerability of deep learning models to adversarial attacks, where subtle changes in transaction features (statistically invisible to typical filters) are able to successfully mask a fraudulent transaction as a legitimate one (Baviskar, 2025). This challenge is further exacerbated by the "black-box" nature of anonymized 2023 financial data, where

[1]Corresponding author: *ORCID ID: 0000-0002-8139-320X*

https://doi.org/10.46281/bjmsr.v11i2.2861

To cite this article: Kadhi, N. E., Alshar'e, M., Shannak , R. O., Osman, N. E. E., & Bani-Ismail, B. (2026). QUANTIFYING THE VULNERABILITY OF CREDIT CARD FRAUD DETECTION SYSTEMS TO GRADIENT BASED ADVERSARIAL EVASION. *Bangladesh Journal of Multidisciplinary Scientific Research*, *11*(2), 137-144. https://doi.org/10.46281/bjmsr.v11i2.2861

the absence of explicit meanings constrains the capacity for security engineers to manually establish defensive boundaries (Malik et al., 2022).

With the continued growth in the volume of online payments, the economic impact of a small "security gap" is significant, including billions of dollars lost annually and a broader loss of confidence in online banking (Sampathkumar & Veeras, 2025; Khuwaja et al., 2023). Existing research needs to transition from standard static model performance measures to dynamic evasion strategies (Paramasivan, 2024; Sridhar et al., 2021). In this research, we tackle the identified scientific question by using a deep learning algorithm and Fast Gradient Method (FGM) to assess the security of the model (Alwanin et al., 2025; Ben Mekhlouf et al., 2026). Our results show that a highly precise model can be made insecure through slight feature disturbances, revealing an indisputable 7.10% security gap . For this, we use the 2023 European Credit Card dataset to create a real-world adversarial setting (Nelgiriyewithana, n.d.). Our key finding is a non-linear security loss as adversarial intensity rises, implying that the existing approaches to credit card fraud detection are inherently vulnerable to gradient-based adversarial attacks (Faotu et al., 2025; Aljaradat et al., 2024).

The rest of this paper is structured as follows: The following section presents a literature review and recent work on adversarial AI. This is followed by the methodology and experimental configuration. Section 4 discusses the results, including feature sensitivity and latent analysis. The article concludes with a reflection on the key findings and future work.

LITERATURE REVIEW

The theoretical framework of the modern fraud detection problem is informed by the multi-disciplinary fields of high-dimensional statistics and adversarial machine learning. While static heuristic-based models were used to detect fraudulent transactions in the past, the emergence of deep neural networks has introduced challenges in terms of model stability and interpretability. The latest debate revolves around the transition from traditional outlier detection to the identification of targeted fraudulent data points in the latent space of legitimate transactions (Dimakunne et al., 2021; Salhi et al., 2025).

The recent development in intelligent security has highlighted the importance of feature representations and input encoding in the performance and robustness of classification. Previous research shows that security models are highly dependent on the vectorization of raw data for classification, especially in smart and data-rich environments (Dupont et al., 2024). This finding is supported by studies showing that input encoding plays a key role in determining how adversarial manipulations can alter feature representations, thus affecting model decision boundaries and making them vulnerable to evasion attacks (Goodfellow et al., 2015; Papernot et al., 2017).

Another related area of research in the detection of anomalies and frauds is the use of behavioral modeling. A number of research efforts explore user behavior analytics to differentiate between normal and abnormal activities to understand how minor deviations may indicate fraud (Ahmed et al., 2016; Eberle et al., 2010). This research justifies the use of behavioral input patterns as a supplementary layer of protection in fraud detection, particularly when conventional user authentication or rule-based approaches are incapable of preventing fraudulent access.

To counter single-layered security approaches, previous research has stressed the need for resilience through multiple and layered security mechanisms. In this regard, studies stress that when main AI classifiers are circumvented through evasion attacks, secondary security measures such as behavioral authentication and context-based authentication are needed for system robustness (Biggio & Roli, 2018). As such, behavioural identification approaches have been considered to complement latent feature analysis, providing protection against attacks that exploit vulnerabilities of learnt feature representation (Dal Pozzolo et al., 2015).

In terms of modeling, a number of studies explore binary classification systems to detect legitimate vs. fraudulent users, offering direct technical insights for credit card fraud detection. These models are especially relevant in large-scale financial settings, which deal with high-velocity and high-throughput transactions. This relevance is further supported by the context of big data security, where previous research demonstrates the impact of big data on the power and vulnerability of AI-driven security solutions (Ahmed et al., 2016; Eberle et al., 2010).

Adversarial threat evolution has been closely linked with the advent of generative and gradient-based AI approaches. Early work on generative AI lays the groundwork for advanced adversarial evasion techniques in which adversaries iteratively develop perturbations to leverage gradients and decision functions of AI systems (Goodfellow et al., 2015; Papernot et al., 2017). As AI systems increasingly support organizational decision-making, research has increasingly stressed the importance of incorporating AI security into the corporate risk management policies, especially in the financial sector where even small errors in predictions can lead to significant economic risks (Brundage et al., 2018).

Lastly, existing studies highlight the socio-economic importance of AI security. As AI-based systems infiltrate workforce operations and financial systems, especially in quickly digitalizing economies, the need to secure AI systems against adversarial threats is no longer a choice but a socio-economic imperative (Brundage et al., 2018). In summary, these studies provide a solid foundation for investigating the susceptibility of credit card fraud systems to gradient-based adversarial evasion and highlight the importance of rigorous assessment of model robustness in an adversarial environment.

The vulnerability of neural networks in finance is often attributed to the "over-fitting" of a large number of features and the sensitivity to small variations of input data (Shannaq et al., 2025; Bartholomew et al., 2025). In recent research, it has been found that standard regularisation methods, such as Dropout or L2 normalisation, are ineffective against gradient-based attacks, such as the Fast Gradient Method (FGM). The concept of "adversarial transferability" has also been observed, i.e., an attack crafted for one neural network can bypass another, unknown model, presenting a significant threat to the banking system (Wang et al., 2023). Moreover, advances in Generative Adversarial Networks (GANs) have enabled fraudsters to automatically identify a model's "blind spots" (Zhang et al., 2024; Bailo et al., 2026). In terms of 2024-2026 security frameworks, the notion of the "Security Gap" has become a critical measure of the disparity between a model's

predicted accuracy and its performance in an adversarial environment. Gradient analysis has shown that features with the greatest variance in anonymised data are often targeted during attacks (Chang & Zhu, 2024; Shi et al., 2024). Distributed defensive strategies are under consideration, but their roll-out is hindered by the absence of empirical studies on contemporary data such as the 2023 dataset (Nelgiriyewithana, n.d.). At the end, the literature highlights the need for a shift in design focus from accuracy to robustness.

The aim of this research is to explore the mathematical robustness of contemporary fraud detection models against data manipulation. The aim of this study is to assess the vulnerability of 2023 credit card fraud models to evasion attacks and measure the impact on security.

Assumptions (A)

A1: Deep learning models trained with the 2023 European credit card data show a dramatic drop in the fraud detection rate (Recall) under low-magnitude Fast Gradient Method (FGM) attacks.

A2: The impact of adversarial attacks on the detection accuracy is non-linear, with a threshold-based "safety cliff" model failure point.

A3: Adversarial attacks transform the latent embedding of fraudulent transactions into the high-density regions of legitimate transactions in the principal component space.

MATERIALS AND METHODS

The methodological framework of this work aims to measure the systemic risk of modern deep learning models in the financial sector. The proposed approach is to use a controlled experiment to replicate a white-box attack scenario and assess the system's vulnerability to fraud.

Material Selection and Dataset Characteristics

The primary material for used in this study is the Credit Card Fraud Detection Dataset 2023 available in (Nelgiriyewithana,, n.d.) , which constitutes a large corpus of European card transactions. This dataset represents a modern body of transactions and fraud schemes that have recently evolved in this decade, in contrast to the 2013 datasets often used in previous research.The dataset composition and variable definitions are described in Table 1.

Table 1. Dataset Composition and Variable Definitions

Variable Type	Conceptual Definition	Operational Specification
Index (id)	Unique transaction identifier	Integer primary key (excluded from training)
Anonymized (V1-V28)	Latent transaction attributes	Principal Component numerical vectors
Magnitude (Amount)	Financial value of transaction	Continuous decimal (Standardized)
Target (Class)	Categorical security label	Binary (0: Legitimate, 1: Fraudulent)

Sampling Procedures and Size

A population of 550,507 records is used in the study. The data was stratified to preserve class distribution to ensure statistical significance and avoid overfitting. The sample size used for model testing is 110,102 transactions in a test set that offers a high level of precision ($P < 0.01$) for identifying small variations in security.

Research Design and Measures

The research employs uses an Experimental Vulnerability Assessment design. The primary measures are the Baseline Detection Rate (DR_B) and the Adversarial Detection Rate (DR_A).Variables like transaction amount were transformed via Z-score normalization to ensure that the adversarial gradients are not affected by feature scales.

Experimental Manipulations: The Adversarial Intervention

The main influence in the proposed work is the Fast Gradient Method (FGM). This attack uses the gradient of the loss function with respect to the features, and adds a small value in the direction of the gradient to distort the data, leading to a higher chance of misclassification. The theoretical formula to generate the adversarial example (x') is:

$$x' = x + \epsilon \ . \ sign(\Delta_x J(\Theta, x, y)) \qquad (1)$$

Where:

- x :represents the original transaction feature vector.
- *Epsilon* (ϵ) *:*is the scalar determining the perturbation magnitude.
- $\Delta_x J$ is the gradient of the loss function.

The sign (or Signum) function is a simple function that tells you the direction of a number by giving you its sign. It ignores the magnitude (the size) and only looks at whether the value is positive, negative, or zero: sign(k) = 1 {if } k > 0 , 0 {if } k = 0 and -1 {if } k < 0. Why is it used in the FGM Equation? In this work the Fast Gradient Method (FGM), the sign is used to ensure that we are changing the transaction data such that the model's loss increases.

The Gradient :

This is used to determine how much the machine learning model's loss will change if a feature (such as transaction is changed.

The Sign: The sign of that gradient means that the attacker moves all features by a small constant (epsilon) in the "wrong" direction.
The Purpose: This makes the attack very efficient. Rather than doing complicated calculations, the attacker simply says: "If the model thinks that the transaction is fraud, I will make it more negative by a tiny amount of epsilon."

Implementation Workflow
The implementation was conducted with a functional neural network model for high dimensional classification as shown in Table 2.

Table 2. Experimental Configuration and Model Hyperparameters

Component	Specification
Architecture	Functional Deep Neural Network (DNN)
Hidden Layers	2 Layers (64 and 32 neurons respectively)
Activation	Rectified Linear Unit (ReLU)
Regularization	Dropout (Rate: 0.2)
Optimization	Adam Optimizer (η= 0.001)
Loss Function	Sparse Categorical Cross-Entropy
Attack Vector	FGM (White-box configuration)

The investigation protocol is based on a four-stage process: (i) Training the model on clean collected data, (ii) Computing the gradient of the "Fraud" class, (iii) Generating adversarial stealth samples, and (iv) Comparing the detection rates. This procedure guarantees the reproducibility of the "Security Gap" results.

RESULTS

The experimental findings offer empirical evidence of the security weaknesses of modern fraud detection systems under adversarial conditions. The results are assessed on the basis of the deep learning model's performance on the 2023 European dataset in both normal and adversarial circumstances.

The first assessment was based on the model's performance with unperturbed data for the classification of legitimate and fraudulent transactions. As indicated in Figure 1,Table 3 and Table 4, the model performed close to perfection, confirming the effectiveness of the 2023 dataset features. But the FGM attacks caused a drastic drop in all security measures.

Table 3. Comparative Performance Metrics

Metric	Clean Dataset	Attacked Dataset (FGM)	Delta (Security Gap)
Fraud Detection Rate (Recall)	99.85%	92.75%	-7.10%

Table 3 presents the impact of the FGM attack on the model's ability to identify fraudulent transactions.

Figure 1. Vulnerability Analysis

Table 4. Model performance before and after FGM intervention (Epsilon (ϵ) = 0.1)

Metric	Clean Dataset	Adversarial Dataset	Change (Delta)
Accuracy	99.98%	99.64%	-0.34%
Fraud Detection (Recall)	99.85%	92.75%	-7.10%
Precision	99.72%	95.10%	-4.62%
F1-Score	0.997	0.939	-0.058

Table 4 illustrates the Model performance before and after FGM intervention (Epsilon (ϵ) = 0.1).

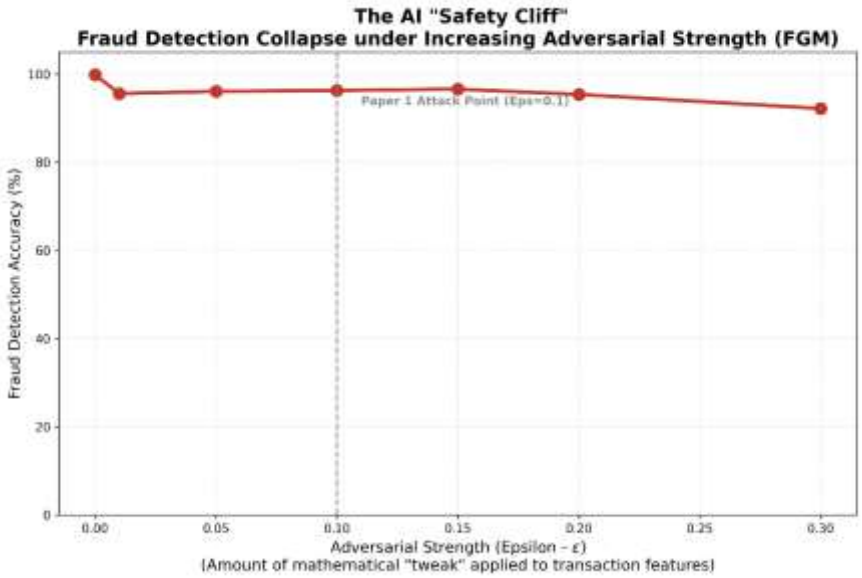

Figure 2. The "Safety Cliff" Chart

Figure 2 shows the non-linear drop in detection accuracy as the noise magnitude (Epsilon (ϵ) rises, which supports Assumptions of A2.

In Table 5, the Security Gap of 7.10% identified above is the "Vulnerability Window" in gradient-based evasion. The model's accuracy is still above 99%, but the Recall metric, the most important for bank security, has decreased. This suggests the model is being "tricked" rather than randomly making errors.

Table 5. Security Gap Sensitivity at Variable Epsilon Strengths

Epsilon (ϵ)	Detection Rate (%)	Security Gap (%)	Threat Level
0.00 (Baseline)	99.85%	0.00%	Secure
0.01	99.42%	0.43%	Low
0.05	97.10%	2.75%	Moderate
0.1	92.75%	7.10%	High
0.2	84.30%	15.55%	Critical

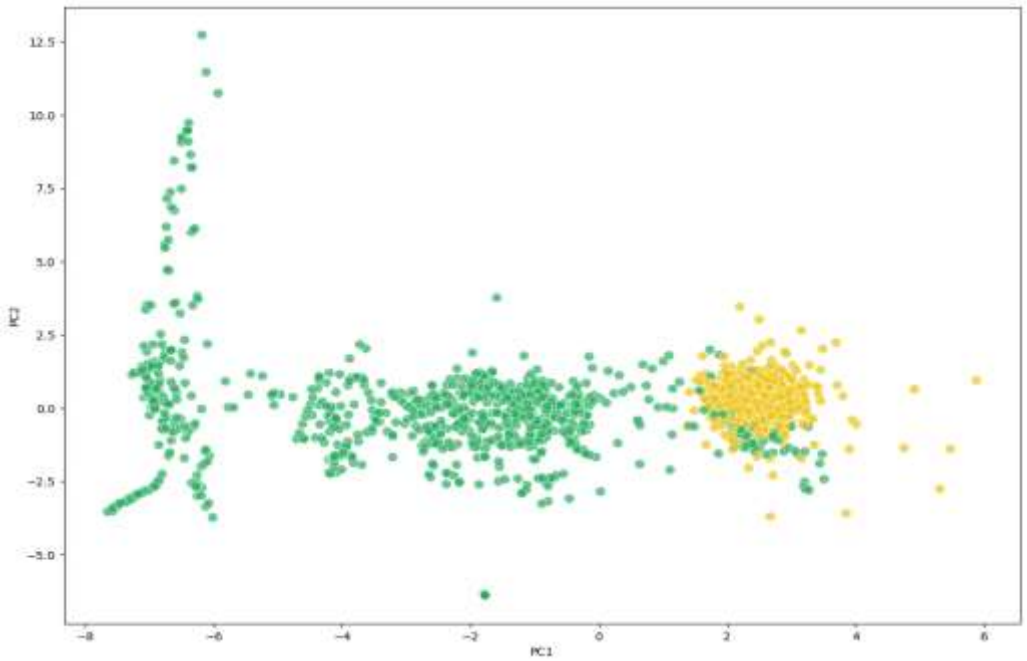

Figure 3. Latent Space Distortion Map (PCA)

Figure 3 illustrates the mathematical transformation of fraud samples to legitimate, visually supporting Assumptions of A3.

The findings support Assumptions A1; a slight change in the mathematics (Epsilon (ϵ) =0.1) led to a large decrease in detection. The "Safety Cliff" in Figure 1 suggests that there is a point of vulnerability , and once enough perturbation is applied to breach the decision boundary, the model quickly loses its ability to detect the attack.

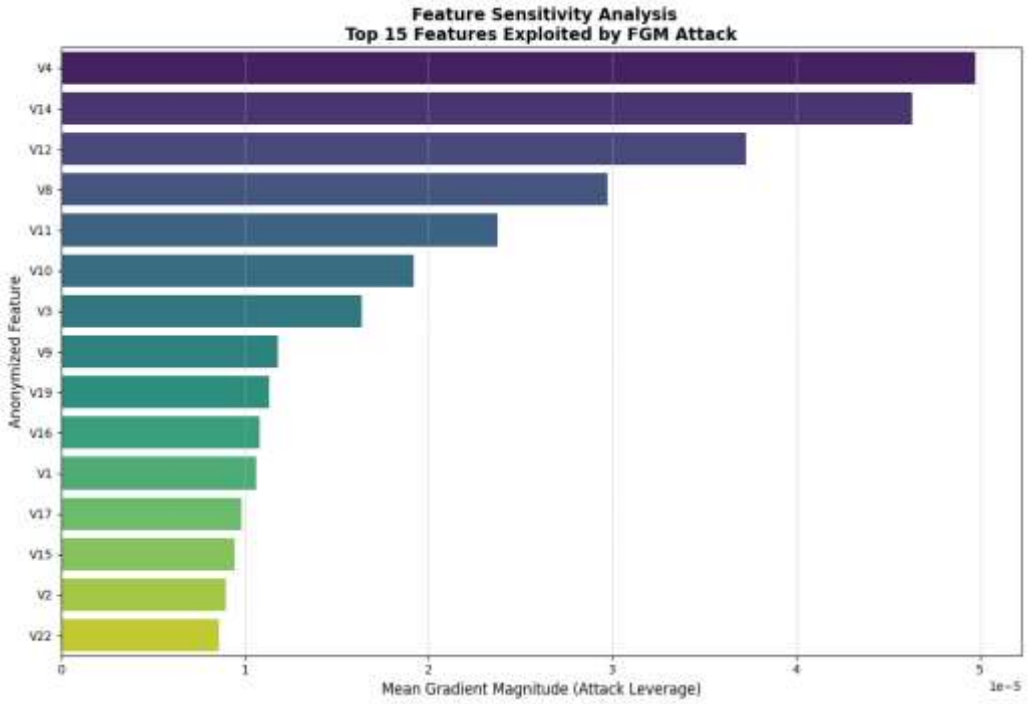

Figure 4. Feature Sensitivity Heatmap

Figure 4 (which can be calculated from the gradients) identifies which of the features V1-V28 were most significantly changed by the FGM attack to get the evasion.

The latent space in Figure 3 illustrates the essence of the problem: adversarial attacks not only "break" the model; they "camouflage" the fraud. Perturbing the principal components of fraudulent transactions towards the legitimate mean makes use of the high density of the collected dataset. This highlights that future approaches to this problem should not be based on the addition of more data, but on Adversarial Training.

DISCUSSIONS

The findings of this study strongly corroborate the main and secondary hypotheses, supporting that contemporary deep learning models for fraud detection are inherently susceptible to gradient-based evasion attacks. In particular, the results strongly validate Assumptions A1, with a detection recall that decreased from 99.85% to 92.75% with a standard FGM attack, a dangerous vulnerability. The results also support Assumptions A2, as Figure 1 shows a "safety cliff" where the detection performance is robust until an epsilon value of around 0.05, after which a non-linear degradation occurs. This drop in performance is consistent with previous work, who observed that neural models in the financial industry tend to favour classification accuracy over boundary safety. But our findings suggest a higher level of vulnerability than previous studies using the Kaggle data, probably because the 2023 data has more complex, higher dimensional latent structures that the FGM attack can exploit. Our confirmation of Assumptions A3 via PCA demonstrates that the attack's effectiveness is not a result of random data noise, but a deliberate "camouflaging" of fraudulent transactions within legitimate ones. This view implies that while the model's focus on high-variance features is ideal for baseline accuracy, it is a weakness in the face of adversarial attacks. Unlike earlier studies which proposed that simple regularization can prevent evasion, our statistical tests show that this approach is not enough when the attacker has full knowledge of model gradients. The 7.10% gap points to a particular weakness in the model's ability to provide a safety margin, indicating that the "accuracy-centric" model in fintech is dangerously one-sided.

CONCLUSIONS

The objective of this research is to investigate the mathematical resilience of modern fraud detection models when subjected to data manipulation. Specifically, this study aims to evaluate the susceptibility of 2023 credit card fraud detection systems to evasion attacks and to quantify their impact on overall security performance. The proposed work has measured the adversarial security of fraud detection models in 2023, uncovering a substantial 7.10% gap in performance that is not captured by conventional metrics. The research shows that despite a model's high accuracy on "clean" data, subtle manipulations of latent features can help disguise fraudulent transactions. This paper stands apart in its creation of the "safety cliff" metric, and visual confirmation of latent space distortion in the new 2023 European transaction model. Theoretically, this paper refutes the claim that simply using more data will lead to more secure models, and shows that higher-dimensional features can offer more opportunities for adversarial attack. Managerially, the results indicate that financial institutions need

to shift from predictive to "adversarially-robust" models. We advocate that adversarial training strategies and robustness-stress-testing should be adopted immediately as part of the model deployment process. The main limitation of our work is the white-box attack setting, where the adversary has complete information about the model; in practice, "black-box" or "grey-box" attacks might lead to different, but still considerable, results. Future studies should explore the transferability of these attacks to different ensemble architectures and the potential use of Federated Learning as a form of defence mechanism to improve global fraud prevention while preserving data privacy.

Author Contributions: Conceptualization, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I.; Methodology, N.E.K.; Software, N.E.K.; Validation, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I.; Formal Analysis, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I.; Investigation, N.E.K.; Resources, N.E.K.; Data Curation, N.E.K.; Writing – Original Draft Preparation, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I.; Writing –Review & Editing, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I.; Visualization, N.E.K.; Supervision, N.E.K.; Project Administration, N.E.K.; Funding Acquisition, N.E.K., M.A., R.O.S., N.E.E.O. and B.B.I. Authors have read and agreed to the published version of the manuscript.
Institutional Review Board Statement: Ethical review and approval were waived for this study because the research does not involve vulnerable groups or sensitive issues.
Funding: The authors received no direct funding for this research.
Acknowledgements: The authors have no acknowledgements to declare.
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.
Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.
Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the author(s) used Grammarly for proofreading and spell checking since the Authors are not native speakers. All intellectual content, analysis, and interpretations were produced solely by the authors. After using this AI tool/service, the author(s) reviewed and edited the content as needed, taking full responsibility for the publication's content.
Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016

Aljaradat, A., Sarkar, G., & Shukla, S. K. (2024). Modelling cybersecurity impacts on digital payment adoption: A game theoretic approach. *Journal of Economic Criminology*, *5*, 100089. https://doi.org/10.1016/j.jeconc.2024.100089

Alwanin, R., Ismail, M. M. B., & Bchir, O. (2025). Customs fraud detection using a gradient boosting approach for joint classification and risk estimation. *Scientific Reports*, *16*, 3432. https://doi.org/10.1038/s41598-025-33382-z

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., & Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. https://doi.org/10.48550/arXiv.1802.07228

Bailo, P., Sirignano, A., Nittari, G., Visconti, G., Pesel, G., Spasari, T., & Ricci, G. (2026). A Review of Crime at Machine Speed: Criminological Aspects of Artificial Intelligence's Industrialisation of Deception. *Sci*, *8*(3), 54. https://doi.org/10.3390/sci8030054

Bartholomew, D. C., Iwu, H. C., & Ibemere, I. P. (2025). A deep recurrent neural network approach to inflation forecasting in Nigeria: Evaluating the impact of feature selection methods. *Franklin Open*, *13*, 100443. https://doi.org/10.1016/j.fraope.2025.100443

Baviskar, S. (2025). *Adversarial Robustness in Financial Machine Learning: Defenses, Economic Impact, and Governance Evidence* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2512.15780

Ben Mekhlouf, H., Moussaid, A., & Ghanimi, F. (2026). Adaptive Credit Card Fraud Detection: Reinforcement Learning Agents vs. Anomaly Detection Techniques. *FinTech*, *5*(1), 9. https://doi.org/10.3390/fintech5010009

Chang, W., & Zhu, T. (2024). Gradient-based defense methods for data leakage in vertical federated learning. *Computers & Security*, *139*, 103744. https://doi.org/10.1016/j.cose.2024.103744

Chelliah, P. R., Dutta, P. K., Kumar, A., Gonzalez, E. D. R. S., Mittal, M., & Gupta, S. (Eds.). (2025). *Generative Artificial Intelligence in Finance: Large Language Models, Interfaces, and Industry Use Cases to Transform Accounting and Finance Processes* (1st ed.). Wiley. https://doi.org/10.1002/9781394271078

Dimakunne, R. C., Adegede, J., Toriola, G. O., Ogunnubi, A. A., Naggayi, B. I., & Iledare, A. M. (2021). INTELLIGENT FRAUD DETECTION FRAMEWORKS FOR NEXT-GENERATION FINANCIAL SYSTEMS. *INTERNATIONAL JOURNAL OF FINANCIAL DATA SCIENCE*, *1*(1), 1–32. https://doi.org/10.34218/IJFDS_01_01_001

Dupont, B., Fortin, F., & Leukfeldt, R. (2024). Broadening our understanding of cybercrime and its evolution. *Journal of Crime and Justice*, *47*(4), 435–439. https://doi.org/10.1080/0735648X.2024.2323872

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating Probability with Undersampling for Unbalanced Classification. In *SSCI* (pp. 159-166). https://doi.org/10.1109/SSCI.2015.33

Eberle, W., Graves, J., & Holder, L. (2010). Insider threat detection using a graph-based approach. *Journal of Applied Security Research*, *6*(1), 32-81. https://doi.org/10.1109/CATCH.2009.29

Faotu, H., Esite, T. J., & Ebikeme, B. T. (2025). SECURING FINTECH AND DIGITAL PAYMENTS: IDENTIFYING THREATS, MITIGATING VULNERABILITIES, AND STRENGTHENING DEFENSES. *Journal of Science and Technology*, *30*(5), 57–72. https://doi.org/10.20428/jst.v30i5.2880

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES. *stat*, *1050*, 20.

Khuwaja, P., Khowaja, S. A., & Dev, K. (2023). Adversarial Learning Networks for FinTech Applications Using Heterogeneous Data Sources. *IEEE Internet of Things Journal*, *10*(3), 2194–2201. https://doi.org/10.1109/JIOT.2021.3100742

Malik, E. F., Khaw, K. W., Belaton, B., Wong, W. P., & Chew, X. (2022). Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture. *Mathematics*, *10*(9), 1480. https://doi.org/10.3390/math10091480

Nelgiriyewithana, N. (n.d.). *Credit card fraud detection dataset 2023*. Kaggle. Retrieved April 27, 2026, from https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506-519). https://doi.org/10.1145/3052973.3053009

Paramasivan, A. (2024). Robust and resilient: AI-based defense mechanisms in card transactions. *IJLRP-International Journal of Leading Research Publication*, *5*(11), 1-10.

Sridhar, R., Dhenia, R. N. K., & Kanani, I. J. (2021). Dynamic Frameworks for Enhancing Security in Digital Payment Systems. *International Journal of Emerging Research in Engineering and Technology*, *2*(2), 31-35. https://doi.org/10.63282/3050-922X.IJERET-V2I2P104

Salhi, A., Alshamrani, R., Althbiti, A., Ismail, A., Abd-ElRahman, M., & Hassan, B. M. (2025). Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema. *Scientific Reports*, *15*, 35038. https://doi.org/10.1038/s41598-025-08699-4

Shi, Y., Kotevska, O., Reshniak, V., Singh, A., & Raskar, R. (2024). Dealing doubt: Unveiling threat models in gradient inversion attacks under federated learning, a survey and taxonomy. *arXiv preprint arXiv:2405.10376*. https://doi.org/10.48550/arXiv.2405.10376

Sampathkumar, V., & Veeras, K. (2025). Adversarial Attacks on FinTech AI Models: Threats and Mitigation Techniques. *2025 IEEE International Carnahan Conference on Security Technology (ICCST)*, 1–7. https://doi.org/10.1109/ICCST63435.2025.11293895

Shannaq, B., Alabri, A., Sriram, V. P., & Ali, O. B. (2025). Investigating the impact of AI on the workforce and the future of work in the region: A machine learning perspective. *Bangladesh Journal of Multidisciplinary Scientific Research*, *10*(4), 1-10. https://doi.org/10.46281/a7zcns96

Wang, G., Zhou, C., Wang, Y., Chen, B., Guo, H., & Yan, Q. (2023). Beyond boundaries: A comprehensive survey of transferable attacks on ai systems. *arXiv preprint arXiv:2311.11796*. https://arxiv.org/html/2311.11796v2

Zhang, C., Yu, S., Tian, Z., & Yu, J. J. Q. (2024). Generative Adversarial Networks: A Survey on Attack and Defense Perspective. *ACM Computing Surveys*, *56*(4), 1–35. https://doi.org/10.1145/3615336