

# Foe for Fraud: Transferable Adversarial Attacks in Credit Card Fraud Detection

Jan Lum Fok<sup>a,b</sup>, Qingwen Zeng<sup>a</sup>, Shiping Chen<sup>c</sup>, Oscar Fawkes<sup>b</sup>, Huaming Chen<sup>a</sup>

<sup>a</sup> The University of Sydney, Australia

<sup>b</sup> Constantinople, Australia

<sup>c</sup> CSIRO, Australia

Jan@cxnpl.com, {qingwen.zeng, huaming.chen}@sydney.edu.au, shiping.chen@data61.csiro.au, OscarF@cxnpl.com

**Abstract**—Credit card fraud detection (CCFD) is a critical application of Machine Learning (ML) in the financial sector, where accurately identifying fraudulent transactions is essential for mitigating financial losses. ML models have demonstrated their effectiveness in fraud detection task, in particular with the tabular dataset. While adversarial attacks have been extensively studied in computer vision and deep learning, their impacts on the ML models, particularly those trained on CCFD tabular datasets, remains largely unexplored. These latent vulnerabilities pose significant threats to the security and stability of the financial industry, especially in high-value transactions where losses could be substantial. To address this gap, in this paper, we present a holistic framework that investigate the robustness of CCFD ML model against adversarial perturbations under different circumstances. Specifically, the gradient-based attack methods are incorporated into the tabular credit card transaction data in both black- and white-box adversarial attacks settings. Our findings confirm that tabular data is also susceptible to subtle perturbations, highlighting the need for heightened awareness among financial technology practitioners regarding ML model security and trustworthiness. Furthermore, the experiments by transferring adversarial samples from gradient-based attack method to non-gradient-based models also verify our findings. Our results demonstrate that such attacks remain effective, emphasizing the necessity of developing robust defenses for CCFD algorithms.

**Index Terms**—Adversarial Attacks, Credit Card Fraud Detection (CCFD), Tabular Data, Transferability, Financial Services

## I. INTRODUCTION

We are witnessing the widespread adoption of machine learning (ML) in financial services, which have revolutionized the areas of fraud detection, risk assessment, and anti-money laundering (AML). To date, these financial services have heavily relied on traditional ML models such as decision trees, logistic regression, and ensemble learning techniques. As a critical application of ML in financial security services, credit card fraud detection (CCFD) is an important use case which enables real-time identification of fraudulent transactions and enhancing fraud prevention strategies. Since financial institutions are now heavily relying on these ML-driven risk assessment, anomaly detection, and fraud detection systems, ensuring their robustness against adversarial threats has become a pressing concern [1].

On the other hand, existing studies have demonstrated vulnerabilities in ML models trained on image data, in particular

by introducing imperceptible perturbations to input data to deceive ML models. These attacks will generally leverage the structured nature of visual features to craft adversarial examples [2]. Many of the works focus on computer vision concerning these adversarial attacks, however it remains underexplored in finance service application such as CCFD which generally utilize tabular datasets. For tabular data, which is commonly used in financial applications, presents unique challenges due to its discrete, heterogeneous feature distributions and domain-specific constraints [3]. This lack of research poses a significant security risk, as financial institutions widely deploy ML-based fraud detection models without fully understanding their vulnerabilities to adversarial manipulation. Statistical reports highlight that credit card fraud caused an estimated loss of £574.2 million in the United Kingdom in 2020 alone [4]. The extent to which established adversarial attack methods can be applied to ML models trained on tabular data for financial decision-making remains unclear, leaving a critical gap in securing ML-based fraud detection systems.

To address this gap, this paper first examines the security of the most commonly used gradient-based ML models in the CCFD domain. We employ gradient-based attack algorithms to generate adversarial samples and investigate their impacts on fraud detection performance. To highlight broader security threats beyond traditional gradient-based architectures and uncover additional attack vectors, we investigate whether the adversarial samples can also mislead a non-gradient-based ML model. This investigation reveals critical vulnerabilities in CCFD models, demonstrating that ML models trained on tabular data are susceptible to both black-box and transfer adversarial attacks. Our findings highlight the urgent need for robust defense mechanisms to enhance the security of ML-based fraud detection systems.

In summary, this paper makes following key contributions:

- We present a holistic framework for transferable adversarial attacks, in particular for CCFD. Our objective is to investigate the underlying issue whereby subtle perturbations to input samples can mislead ML models in the context of fraud detection. To this end, we evaluate a gradient-based model, revealing that tabular data may serve as a susceptible attack surface for adversaries. (Sec.III)

- We perform a comprehensive evaluation of the proposed framework using real-world credit card transaction datasets. Furthermore, we evaluate the effectiveness of the adversarial samples on unseen ML models, which shares entirely different model architecture and training process. In this paper, we consider a non-gradient-based ML model to assess the transferability of the adversarial attacks. With the collected results based on the attack success rate, our work demonstrates that adversarial samples generated in a white-box setting can successfully transfer to a structurally different model while maintaining high attack efficacy. (Sec.IV)
- Through the framework, we provide empirical insights about the security vulnerabilities of CCFD models, revealing their susceptibility to adversarial attacks, including black-box and transfer attacks. Our findings highlight the urgent need to strengthen the security of financial fraud detection systems and develop more robust defenses against adversarial threats. (Sec.V)

The remainder of this paper is structured as follows: Section II reviews related work. Section III introduces the techniques used in our proposed framework. Section IV details the experimental setup. Section V presents the experimental results and provides an in-depth analysis of their implications. Finally, Section VI concludes the paper and outlines potential future research directions.

## II. LITERATURE REVIEW

Credit card fraud detection (CCFD) is a prominent application of machine learning (ML) in financial security, where traditional rule-based systems have been increasingly replaced by supervised models such as logistic regression, Naive Bayes, KNN, random forests, and SVM for improved detection performance [5]–[7]. Deep neural networks have also been adopted to further enhance detection accuracy [8]. However, these models mainly focus on classification performance, while their robustness against adversarial threats remains underexplored [9]. Adversarial attacks—crafted perturbations designed to mislead model predictions—have been extensively studied in image-based domains using methods like the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner attacks [10]–[12]. However their adaptation to tabular data, especially in CCFD, poses unique challenges due to the discrete and heterogeneous nature of features [3]. Although some works have examined adversarial risks in CCFD, they largely concentrate on deep learning under white-box assumptions [13], [14], and penetration testing appears insufficient against such threats [15]. Meanwhile, commonly deployed traditional models like logistic regression and random forest—still favored in fraud detection for their simplicity and efficiency—have received little attention in this context [16]. Moreover, the feasibility of applying transferable adversarial attacks in black-box settings for financial services remains unvalidated, forming the core research gap this work aims to address.

## III. PROPOSED FRAMEWORK

The objective of this work is to investigate the security of machine learning (ML) models in credit card fraud detection (CCFD) within the financial sector. Given that this field remains relatively nascent, existing research on its security vulnerabilities is still limited. CCFD based on ML algorithms is typically formulated as a binary classification problem, where transactions are classified as either legitimate or fraudulent (e.g., credit card skimming) based on extracted features such as transaction time, transaction amount, frequency, and type. ML models will make these classifications using a predefined decision threshold. For adversarial threats, if an attacker obtains knowledge of a model’s decision boundary or the inherent architecture and parameters, they can exploit this information by introducing small, strategically crafted perturbations along the gradient direction. These adversarial perturbations can alter fraud probabilities just enough to shift a fraudulent transaction below the classification threshold, causing a misclassification as a legitimate transaction.

In this section, we will first present the overall framework to generate the adversarial samples following by the design of white-box and black-box adversarial attacks. As shown in Fig. 1, we first introduce an adversarial attack method, which is the Fast Gradient Sign Method (FGSM), to generate the adversarial samples for a gradient-based ML model. The goal is to evaluate whether adversarial attacks remain effective in the context of CCFD, particularly for tabular dataset. To further investigate potential attack vectors in this domain, we extract adversarial samples that successfully bypass detection in the first stage and utilize them as attack inputs against a non-gradient-based ML model. This step examines whether transferable attacks can compromise CCFD systems even when the target model’s internal structure and parameters are entirely unknown.

Through this framework, we aim to uncover vulnerabilities in both ML algorithms and the underlying tabular data used in fraud detection. Our findings emphasize the necessity of integrating security measures into the development and deployment of ML-based fraud detection models in the financial sector to mitigate adversarial threats.

### A. Machine learning models in our Framework

1) *Gradient-based ML*: Most of the gradient-based ML models rely on gradient descent and its variants to optimize the model parameters. For CCFD, logistic regression (LR) is a widely used gradient-based ML algorithm that estimates the probability of an event occurring based on input features [17]–[19]. Unlike linear regression, which models a continuous output, LR applies the logistic function to map predictions to probabilities within the range of 0 and 1 [20]. The logistic function is defined as:

$$\text{Logit}(\pi) = \frac{1}{1 + \exp(-\pi)} \quad (1)$$

where  $\pi$  represents the linear combination of input features and model parameters.

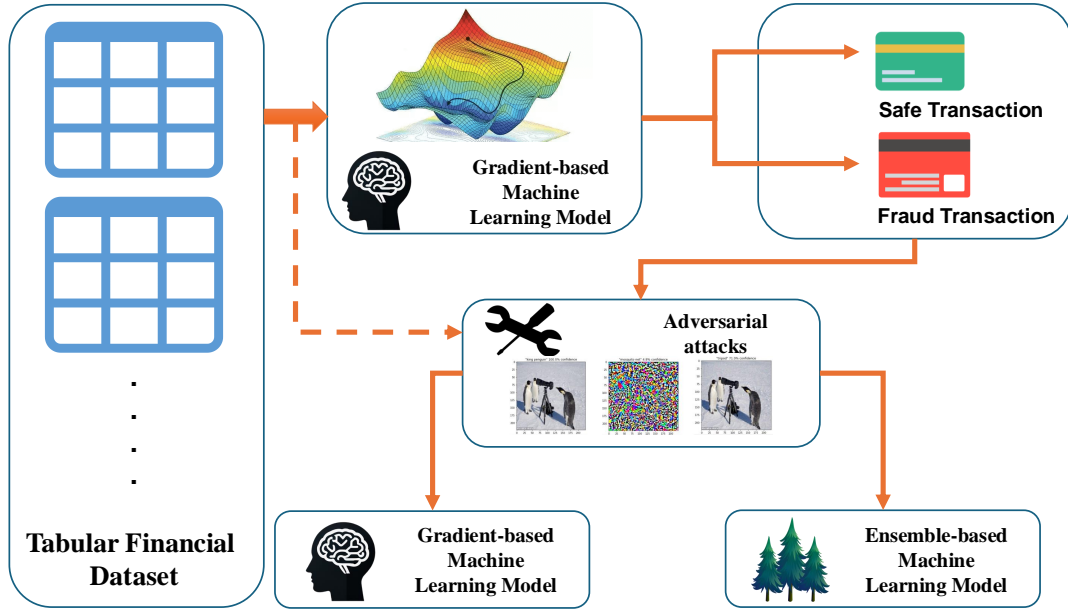


Fig. 1. The overall framework of FGSM-based adversarial attack on credit card fraud detection models

The model parameters are optimized using maximum likelihood estimation (MLE) to minimize classification error. Once trained, LR predicts probabilities, and classification is performed by assigning labels based on a threshold (e.g.,  $\pi \geq 0.5$  is classified as class 1, otherwise class 0) [20]. The sigmoid function, which characterizes LR, ensures that predictions are bounded within the probability range.

2) *Non-gradient-based ML*: A different type of ML model will be non-gradient-based models, which have entirely different architectures and training processes when compared to their gradient-based counterparts. These models generally select features and categorizes based on criteria like Gini impurity or entropy information. An example model is random forest (RF), which is an ensemble learning algorithm designed for classification and regression tasks. It constructs multiple decision trees, each trained on randomly selected subsets of the dataset and feature space [21]. The randomness in RF is introduced through a process known as bootstrap aggregation (bagging), where different samples of the training data are used to build individual trees.

For classification tasks, RF determines the final prediction by applying a majority voting scheme across all decision trees. In regression tasks, it computes the average prediction from all trees. The introduction of randomness at both the data sampling and feature selection levels enhances model generalization and mitigates the risk of overfitting [22].

#### B. Adversarial attack techniques in our framework

While numerous adversarial attack techniques exist in the literature, this work focuses on transferable adversarial attacks in the context of CCFD. To achieve this, we employ the FGSM as the primary attack technique. FGSM, introduced by Goodfellow et al. [10], demonstrates the vulnerability of ML

models to adversarial perturbations. It generates adversarial examples by computing the gradient of the loss function with respect to the input data and applying a small perturbation in the direction of the gradient:

$$\eta = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (2)$$

where  $\epsilon$  is a small perturbation factor,  $\theta$  represents the model parameters,  $x$  is the input,  $y$  is the target label, and  $J(\theta, x, y)$  is the loss function.

The perturbed adversarial example is computed as:

$$x^* = x + \eta = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (3)$$

where  $x^*$  represents the generated adversarial sample. By leveraging the model's gradient, FGSM efficiently alters input features to deceive the classifier while maintaining minimal perceptual changes.

## IV. EXPERIMENTS

### A. Dataset

We use a publicly available credit card transaction dataset from Kaggle [23], containing 284,807 transactions from European cardholders over two days in 2013. The dataset is highly imbalanced, with only 0.17% (492) labeled as fraudulent, reflecting real-world conditions. To preserve confidentiality, features were anonymized using PCA, resulting in variables V1–V28, while ‘Time’ and ‘Amount’ were retained. To address class imbalance, we apply the SMOTE technique [24] to oversample fraudulent cases. The dataset is split into 80% training and 20% testing using stratified sampling to maintain class distribution.

### B. Evaluation Metrics

To comprehensively evaluate model performance on the imbalanced credit card fraud detection (CCFD) task, we report standard metrics including **Accuracy**, **Precision**, and **Recall**. Accuracy reflects overall prediction correctness, while Precision and Recall capture the model’s ability to correctly identify fraud cases and minimize false negatives. In addition, we introduce the Transferability Rate to assess how well adversarial samples generated from one model can mislead another:

$$\text{Transferability Rate} = \frac{\text{Misclassified Samples in Target Model}}{\text{Total Adversarial Samples}}$$

This metric is crucial for evaluating security risks in multi-classifier fraud detection scenarios.

### C. Experimental Design

We adopt logistic regression (LR) as the baseline model due to its high interpretability and relevance in financial fraud detection. Its reliance on gradient-based optimization makes it vulnerable to adversarial attacks such as the Fast Gradient Sign Method (FGSM). The model is implemented using the scikit-learn library [25] and serves as the primary attack target.

To evaluate model robustness, FGSM attacks are applied using the Adversarial Robustness Toolbox (ART) [26] in a white-box setting with full access to model gradients. Adversarial perturbations target correctly classified fraudulent transactions in the test set, aiming to flip them to non-fraudulent labels through minimal input changes.

To assess transferability, we test whether adversarial examples crafted for LR can mislead a non-gradient-based model—random forest (RF)—trained on the same dataset with an 80/20 stratified split. This transfer-based black-box attack reveals cross-model vulnerabilities and highlights the broader risks of adversarial examples across heterogeneous fraud detection systems.

## V. RESULTS AND ANALYSIS

We present Fast Gradient Sign Method (FGSM) attack results on a Logistic Regression (LR) model for credit card fraud detection (CCFD), analyzing recall degradation under varying  $\epsilon$  and the transferability of adversarial samples to a non-gradient-based model (Random Forest).

### A. Baseline Model Performance

To establish a reference point for evaluating the impact of adversarial attacks, we first trained and evaluated a LR model on the Kaggle CCFD dataset. The evaluation metrics for the baseline model are as follows in Table I :

TABLE I  
MODEL EVALUATION METRICS

| Metric    | Value |
|-----------|-------|
| Accuracy  | 0.99  |
| Precision | 0.17  |
| Recall    | 0.92  |

The high recall value indicates that the model effectively detects fraudulent transactions under normal conditions. However, the relatively low precision suggests that the model produces a significant number of false positives, a trade-off that is acceptable given the priority of recall in fraud detection.

### B. Impact of FGSM on Logistic Regression

To assess the vulnerability of the LR model to adversarial attacks, we generated adversarial samples using FGSM, targeting fraudulent transactions that were correctly classified in the test set. The  $\epsilon$  value was set to 2.2, and the generated adversarial samples were used in place of their benign counterparts. The model’s performance on the adversarial test set is shown below in Table II:

TABLE II  
MODEL EVALUATION METRICS

| Metric    | Value |
|-----------|-------|
| Accuracy  | 0.99  |
| Precision | 0.11  |
| Recall    | 0.56  |

Compared to the original recall of 0.92, the recall dropped significantly, indicating that nearly 40% of fraudulent transactions were misclassified as non-fraudulent. This significant reduction in fraud detection capability underscores the security risk posed by adversarial attacks, highlighting the need for robust defense mechanisms in CCFD systems.

### C. Effect of Epsilon on Model Robustness

To analyze how different adversarial perturbation magnitudes affect model performance, we varied the  $\epsilon$  value and observed its impact on recall. The results are plotted in Fig. 2.

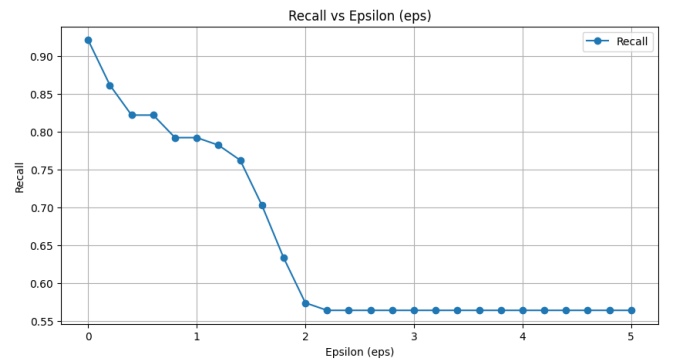


Fig. 2. Effect of increasing adversarial perturbations ( $\epsilon$ ) on model recall, showing degradation trends and eventual stabilization.

From Fig. 2, we observe the following trends:

- At  $\epsilon = 0$ , the recall is 0.92, indicating correct detection of 92% of fraudulent transactions without adversarial perturbation.
- As  $\epsilon$  increases, recall steadily decreases, showing the increasing effectiveness of the FGSM attack.
- When  $\epsilon = 2.2$ , recall drops to 0.56, meaning that nearly half of the fraudulent transactions are misclassified.

- For  $\epsilon > 2.2$ , recall stabilizes at 0.56, suggesting that increasing perturbation magnitude beyond this point does not further degrade model performance.

This saturation effect implies that FGSM has a maximum attack effectiveness threshold, beyond which further perturbations do not introduce additional vulnerability. One possible explanation is that as perturbations increase, they may push fraudulent samples too far from their original distribution, making them unrealistic and potentially easier to detect through anomaly detection mechanisms. Additionally, excessive perturbations could move samples beyond the model's decision boundary in a way that no longer affects classification, leading to diminishing returns in attack effectiveness.

#### D. Analysis of Feature Perturbations

To better understand the effect of adversarial perturbations on individual features, we analyzed the original and adversarial values of a selected successfully perturbed fraudulent transaction. Table III presents the perturbation details.

TABLE III  
ADVERSARIAL PERTURBATION ANALYSIS FOR A SUCCESSFULLY MISCLASSIFIED FRAUDULENT SAMPLE.

| Feature | Before (Original) | After (Adversarial) | Perturbation |
|---------|-------------------|---------------------|--------------|
| Time    | 48884.0           | 48886.2             | 2.2          |
| V1      | -2.13905056       | -4.3390503          | -2.19999974  |
| V2      | 1.39436766        | -0.8056323          | -2.19999996  |
| V3      | -0.612034895      | -2.8120348          | -2.19999991  |
| V4      | 1.04932706        | -1.1506729          | -2.19999996  |
| V5      | -1.16210191       | -3.3621018          | -2.19999989  |
| V6      | -0.768219363      | 1.4317807           | 2.20000006   |
| V7      | -1.99723740       | 0.20276256          | 2.19999996   |
| V8      | 0.574996543       | 2.7749965           | 2.19999996   |
| V9      | -0.980831776      | 1.2191682           | 2.19999998   |
| V10     | -2.49561925       | -0.2956193          | 2.19999995   |
| V11     | 2.55558915        | 0.3555892           | -2.19999995  |
| V12     | -3.53043627       | -1.3304362          | 2.20000007   |
| V13     | -1.01623382       | 1.1837661           | 2.19999992   |
| V14     | -3.45519658       | -1.2551966          | 2.19999998   |
| V15     | -0.056363864      | 2.1436367           | 2.20000009   |
| V16     | -2.46773703       | -0.26773703         | 2.2          |
| V17     | -7.14032597       | -4.940326           | 2.19999997   |
| V18     | -1.27128002       | 0.92871994          | 2.19999996   |
| V19     | -0.00172192354    | -2.201722           | -2.20000008  |
| V20     | 0.0254265126      | 2.2254264           | 2.19999989   |
| V21     | 0.696954881       | -1.5030451          | -2.19999998  |
| V22     | 0.740003045       | -1.4599969          | -2.19999995  |
| V23     | -0.155115249      | -2.3551152          | -2.19999995  |
| V24     | -0.0506074461     | -2.2506075          | -2.20000005  |
| V25     | 0.268368293       | -1.9316317          | -2.19999999  |
| V26     | -0.469432841      | 1.7305672           | 2.20000004   |
| V27     | -0.405813768      | -2.6058137          | -2.19999993  |
| V28     | -0.152170847      | -2.352171           | -2.20000015  |
| Amount  | 19.73             | 17.53               | -2.2         |

Key observations from Table III reveal the behavior of the model under adversarial manipulation:

- The PCA-transformed features (V1–V28) exhibit consistent perturbations of approximately  $\pm 2.2$ , which are sufficient to mislead the logistic regression model. Since these features are derived from principal component analysis, even small changes can significantly alter the encoded representations.

- Named features such as Time and Amount remain nearly unchanged, indicating that adversarial perturbations target abstract feature spaces rather than raw transaction details.
- These findings highlight the model's sensitivity to transformed inputs and underscore the importance of considering adversarial robustness during the preprocessing and feature engineering stages in fraud detection pipelines.

#### E. Transferability of Adversarial Samples

To further evaluate the impact of adversarial samples, we test their ability to transfer to another classification model. Specifically, we assess whether the adversarial samples generated using the FGSM attack on the logistic regression model can also deceive a non-gradient-based model such as Random Forest (RF).

1) *Baseline RF Model Performance:* An RF classifier was trained on the same dataset as the LR model, following the same 80/20 train-test split. The baseline performance of RF on benign test samples is reported below in Table IV:

TABLE IV  
MODEL EVALUATION METRICS

| Metric    | Value |
|-----------|-------|
| Accuracy  | 1.00  |
| Precision | 1.00  |
| Recall    | 0.95  |

TABLE V  
CONFUSION MATRIX FOR BASELINE RANDOM FOREST MODEL BEFORE ADVERSARIAL ATTACK.

| Actual    | Predicted: Non-fraud | Predicted: Fraud |
|-----------|----------------------|------------------|
| Non-fraud | 56861                | 5                |
| Fraud     | 0                    | 96               |

As shown in Table V, RF achieves near-perfect performance in fraud detection under normal conditions.

2) *Adversarial Sample Transferability Experiment:* To examine the robustness of RF against adversarial samples, we applied the adversarial test set generated using FGSM on the LR model to the trained RF model. The results are as follows in Table VI:

TABLE VI  
ADVERSARIAL ATTACK RESULTS

| Metric                       | Value |
|------------------------------|-------|
| Successful attacks           | 34    |
| Failed attacks               | 2     |
| Transferability success rate | 94%   |

The high success rate (94%) indicates that adversarial perturbations crafted for a gradient-based model like LR can still significantly degrade the performance of a non-gradient-based model such as RF. This demonstrates that adversarial attacks are not limited to models that rely on gradients, but can also transfer across different types of classifiers.

3) *Analysis of Transferability*: Several factors contribute to the high transferability of adversarial samples:

- Feature-space perturbations: Even though RF does not use gradients, it relies on feature importance for classification. Perturbations introduced by FGSM may shift the decision boundary for critical features, leading to misclassification.
- Shared decision boundaries: Both LR and RF models were trained on the same dataset, meaning they may learn similar decision boundaries. This increases the likelihood that adversarial perturbations effective for LR also impact RF.
- Structural limitations of ensemble models: While RF benefits from bagging and feature randomness, it remains susceptible to adversarial perturbations that systematically alter the input distribution, leading to incorrect classifications.

These results highlight the broader security risks of adversarial attacks in real-world fraud detection systems. Even when an organization deploys multiple classifiers to improve robustness, transferability can still expose all models to adversarial threats. Future research should explore adversarial training techniques and other defense mechanisms to mitigate these risks.

## VI. CONCLUSION

In this paper, we have examined the vulnerability of credit card fraud detection (CCFD) models to adversarial attacks, demonstrating that adversarial samples generated via Fast Gradient Sign Method (FGSM) significantly degrade model performance. In particular, logistic regression model's recall dropped from 92% to 56%, and these perturbations exhibited a 94% transferability rate to a non-gradient-based Random Forest model, underscoring the broader security risks posed by adversarial attacks in financial applications.

While various adversarial defense mechanisms have been proposed—such as adversarial training, feature regularization, and ensemble defenses [10]—most of these methods have been developed for non-tabular datasets. Gradient obfuscation methods aim to prevent attackers from exploiting model gradients [27], while input preprocessing techniques such as noise filtering and data compression attempt to mitigate the impact of adversarial perturbations [28]. Their applicability to CCFD models remains largely unexplored. Future work should focus on adapting adversarial defenses for machine learning (ML) models on tabular datasets, evaluating their effectiveness across diverse financial datasets, and exploring hybrid defense strategies that integrate multiple mitigation techniques.

In conclusion, adversarial attacks present a systemic security risk to ML-based fraud detection, extending beyond gradient-based models. Developing robust, domain-specific adversarial defenses is crucial to ensuring the trustworthiness and resilience of financial fraud detection systems.

## ACKNOWLEDGEMENT

The first author conducted this work at Constantinople Operating Company Pty Ltd (Constantinople), as part of a funded research collaboration with The University of Sydney. The authors would like to thank Constantinople for their support of this work.

## REFERENCES

- [1] D. Lunghi, A. Simitsis, O. Caelen, and G. Bontempi, "Adversarial learning in real-world fraud detection: Challenges and perspectives," *arXiv*, vol. 2307.01390, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2307.01390>
- [2] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.
- [3] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: Application to fraud detection and imbalanced data," *arXiv preprint arXiv:2101.08030*, 2021.
- [4] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 2018, pp. 1–6.
- [5] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022.
- [6] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection - machine learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019, pp. 1–5.
- [7] S. Kumar, V. K. Gunjan, M. D. Ansari, and R. Pathak, "Credit card fraud detection using support vector machine," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*. Springer, 2022, pp. 27–37.
- [8] R. Asha and S. K. KR, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35–41, 2021.
- [9] F. V. Jedrzejewski, L. Thode, J. Fischbach, T. Gorschek, D. Mendez, and N. Lavesson, "Adversarial machine learning in industry: A systematic literature review," *Computers & Security*, p. 103988, 2024.
- [10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: <https://arxiv.org/abs/1412.6572>
- [11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: <https://arxiv.org/abs/1706.06083>
- [12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017. [Online]. Available: <https://arxiv.org/abs/1608.04644>
- [13] A. Agarwal and N. K. Ratha, "Black-box adversarial entry in finance through credit card fraud detection," in *CIKM Workshops*, 2021.
- [14] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi, "Adversarial support vector machine learning," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 1059–1067.
- [15] D. Lunghi, A. Simitsis, and G. Bontempi, "Assessing adversarial attacks in real-world fraud detection," in *2024 IEEE International Conference on Web Services (ICWS)*. IEEE, 2024, pp. 27–34.
- [16] P. Tiwari, S. Mehta, N. Sakhuja, J. Kumar, and A. K. Singh, "Credit card fraud detection using machine learning: a study," *arXiv preprint arXiv:2108.10005*, 2021.
- [17] T. Wang and Y. Zhao, "Credit card fraud detection using logistic regression," in *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 2022, pp. 301–305.
- [18] A. Mahajan, V. S. Baghel, and R. Jayaraman, "Credit card fraud detection using logistic regression with imbalanced dataset," in *2023 10th international conference on computing for sustainable global development (INDIACom)*. IEEE, 2023, pp. 339–342.
- [19] M. V. Krishna and J. Praveenchandar, "Comparative analysis of credit card fraud detection using logistic regression with random forest towards an increase in accuracy of prediction," in *2022 International Conference on Edge Computing and Applications (ICECAA)*. IEEE, 2022, pp. 1097–1101.

- [20] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
- [21] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
- [22] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, pp. 31–39, 2017.
- [23] M. L. G. ULB, "Credit card fraud detection," n.d. [Online]. Available: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data>
- [24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002. [Online]. Available: <http://dx.doi.org/10.1613/jair.953>
- [25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>
- [26] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.0.0," 2019. [Online]. Available: <https://arxiv.org/abs/1807.01069>
- [27] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint*, vol. arXiv:1705.07204, 2017.
- [28] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of jpg compression on adversarial images," *arXiv preprint*, vol. arXiv:1608.00853, 2016.