International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

A Framework for Adversarial Attack Detection in Digital Finance

Gnana Prakash G
B.M.S. College for Women
gnanaprakashg@bmscw.edu.in

Naga Ganesh B J
B.M.S. College for Women
nagaganesh@bmscw.edu.in

Sudheendra M
B.M.S. College for Women
sudheendra@bmscw.edu.in

Divyashree B Rao
B.M.S. College for Women
divyashreebrao@bmscw.edu.in

Abstract

Adversarial attacks threaten the robustness and reliability of machine learning (ML) models in digital finance by intentionally creating inputs that cause classifiers to make errors. The vulnerability of ML models in digital finance to adversarial attacks compromises mission-critical applications such as fraud detection, algorithmic trading, and credit risk assessment. This research presents a robust detection framework developed specifically for digital finance, combining several defense mechanisms and new feature extraction methods to enhance the detection accuracy and efficiency of adversarial attacks. The framework uses a combination of anomaly detection, statistical analysis, and adaptive learning to detect evasion and poisoning attacks on different ML architectures employed in fintech. Experimental evaluations using standard benchmark datasets show that the developed approach yields higher detection rates than existing approaches while maintaining lower false-positive rates and minimal computational overhead required for high-frequency financial operations. The proposed framework has potential to strengthen the security posture of ML-driven systems in real-world digital finance deployments.

Keywords: Adversarial Machine Learning, Digital Finance, FinTech Security, Fraud Detection, Evasion Attacks, Poisoning Attacks, Anomaly Detection.

I. Introduction

Machine learning models have grown ubiquitous in the digital finance ecosystem, supporting key applications like transaction fraud detection, credit scoring, algorithmic trading, and automated customer service.

However, the emergence of adversarial machine learning (AML) has raised concerns about the growing security risks in FinTech: attackers with financial motives will deliberately alter input data to induce model mispredictions or evade security measures [1]. These adversarial attacks, including evasion and poisoning attacks, pose threats to the trustfulness, regulatory compliance, and usefulness of AI-driven financial systems.

The development of robust detection methods is essential to mitigate these threats and protect digital financial assets. Many of the current defenses focus on detecting specific types of attacks or image-based model architectures, which limits their applicability in tabular financial environments [2]. The present study develops a comprehensive adversarial detection framework specifically designed for digital finance

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

that integrates multiple detection strategies. Through the use of anomaly detection and adaptive learning, the framework can effectively detect a wide range of attacks without interfering with normal financial activities. 2 The contributions of the present study are as follows:

• A multi-layered detection framework incorporating statistical features, model uncertainty, and input perturbation analysis tailored to financial data streams.

• The use of adaptive thresholds based on run-time data distributions to enhance detection robustness in highly volatile financial markets.

• Evaluation of the framework on multiple tabular datasets and attack vectors, demonstrating enhanced detection accuracy over the baseline approaches.

II. Background and Related Work

A. Types of Adversarial ML Attacks in Finance

Adversarial attacks in digital finance fall into one of two categories:

- **Poisoning Attacks:** Attacks that introduce malicious samples (e.g., fraudulent transactions appearing as legitimate) into the training data, leading to the model learning incorrect patterns. This results in reduced performance or backdoored behavior post-deployment, allowing subsequent fraud to go unnoticed [3].
- **Evasion Attacks:** Attacks that occur during inference, in which the attacker creates inputs with minor perturbations (e.g., slightly changes the amount of a transaction or time-related metadata) that deceive the model without changing the actual nature of the activity [1][2].

The methods used to create such adversarial samples include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and optimization-based methods developed specifically for tabular and discrete financial data [1][2].

B. Detection Approaches

Existing AML detection techniques include

- **Statistical Methods:** Detect distributional shifts or anomalies in feature space, typically used in traditional fraud detection.
- **Model uncertainty metrics:** Confidence scores or Bayesian uncertainty are used to flag suspicious financial inputs [4].
- **Input Transformation:** Defend by preprocessing inputs to normalize data and remove perturbations.
- **Ensemble and Hybrid Methods:** Combine multiple detection mechanisms for improved accuracy [5].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Although considerable progress has been made in this area, limitations still exist regarding detection accuracy, false-positive rates (which may prevent legitimate customer transactions), and computational costs (which add delay in high-speed trading), particularly when adaptive adversarial strategies are employed [2][4].

III. Proposed Digital Finance Detection Framework

A. Framework Architecture

The framework is designed to integrate seamlessly into financial transaction pipelines and consists of the following components.

1. **Feature Extraction Module:** Extracts multidimensional features from model outputs related to transaction data, including statistical properties, input perturbations, and uncertainty metrics.
2. **Anomaly Detection Engine:** Utilizes clustering and density estimation (such as Isolation Forest or Local Outlier Factor) to identify deviations from normal financial behavior patterns.
3. **Adaptive Thresholding Unit:** Learns dynamic detection thresholds using real-time feedback to minimize false positives and automatically adjusts to market volatility or seasonal transaction spikes.
4. **Decision Aggregator:** Combines outputs from multiple detectors using weighted voting to determine whether an input is adversarial prior to approving or denying a transaction.

B. Feature Extraction Details

The feature extraction module generates a comprehensive representation of each transaction as follows:

- **Statistical Features:** Higher-order statistics (skewness and kurtosis) of input embeddings and intermediate layer activation tracking deviations in financial profiles.
- **Perturbation Analysis:** Estimations of input noise levels and gradient-based perturbation metrics specific to tabular and time-series financial data.
- **Uncertainty Quantification:** Confidence scores from softmax layers, entropy, and Bayesian posterior variance to measure model hesitation in borderline transactions.

This comprehensive set of features allows for detection of subtle adversarial manipulations that cannot be identified by simpler fraud rules [1][3].

C. Adaptive Thresholding

Fixed detection thresholds are susceptible to variations in financial input distributions due to natural changes and increasingly sophisticated attacks. By utilizing a reinforcement learning agent that monitors detection outcomes and historical transaction trends, the system dynamically adjusts the thresholds to strike a balance between security sensitivity and transactional specificity [5].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

IV. Experimental Evaluation

A. Datasets and Attacks

To simulate digital finance environments, experiments were conducted on standard tabular benchmark datasets:

- Credit Card Fraud Detection dataset (Kaggle)
- PaySim mobile money simulator
- IEEE-CIS Fraud Detection dataset

The attack methods included FGSM and PGD adapted for tabular data, alongside targeted poisoning injections [1][2].

B. Baseline Comparisons

The framework was benchmarked against state-of-the-art FinTech detection methods, including:

- Feature squeezing
- Statistical anomaly tests
- Uncertainty-based detectors
- Ensemble defense mechanisms [4][5]

C. Metrics

The detection accuracy, false-positive rate (FPR), and computational overhead (latency) were evaluated to ensure the feasibility of real-time deployment.

D. Results Summary

Metric	Proposed Framework	Baseline Average
Detection Accuracy	95-98%	88-92%
False Positive Rate	<3%	5-8%
Computational Overhead	Low	Medium-High

Table 1: Performance Comparison of Proposed Framework vs Baseline methods

Key findings:

- The proposed framework achieved detection accuracies of 95–98% across various financial attack vectors.
- The false-positive rate was reduced to under 3%, outperforming the baselines and ensuring that legitimate transactions were rarely blocked.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

- Adaptive thresholding significantly improved robustness against varying attack intensities and simulated-market shifts.
- The computational cost was maintained within acceptable limits for real-time FinTech deployment scenarios.

V. Discussion

The multilayered approach harnesses diverse feature perspectives, allowing improved generalization to unseen adversarial tactics targeting financial models. Adaptive thresholding addresses the classic challenge of static detectors in volatile finance sectors, enhancing their real-world feasibility.

However, certain white-box attacks with adaptive financial adversaries require further reinforcement using continual learning strategies. The design also prioritizes modularity for flexible integration into existing banking infrastructure.

The framework demonstrates particular strengths in the following areas:

- Handling imbalanced financial datasets common in fraud detection
- Maintaining low latency suitable for high-frequency trading environments
- Adapting to seasonal and market-driven distribution shifts
- Reducing false positives that impact customer experience

VI. Challenges and Future Directions

Several areas warrant further investigation.

- **GAN-based Attacks:** Extending detection to Generative Adversarial Network (GAN)-based sophisticated attacks designed to synthesize highly realistic fraudulent financial profiles.
- **Privacy-Preserving Detection:** Developing privacy-preserving detection mechanisms (e.g., via Federated Learning) suitable for collaborative threat intelligence sharing between financial institutions [6].
- **Interpretability:** Improving the interpretability of detection decisions to aid in regulatory compliance, auditing, and forensic analysis.
- **Continual Learning:** Incorporating automated feedback loops for continual model retraining in evolving financial threat landscapes.
- **Cross-domain Generalization:** Testing framework effectiveness across diverse financial domains (banking, insurance, investment).

VII. Conclusion

Adversarial attacks critically endanger the reliability of ML systems in the digital finance sector. This study presents a robust detection framework that integrates multiple complementary techniques to achieve strong adversarial input identification tailored for FinTech applications. By maintaining low false alarms

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

and a manageable computational overhead, adaptive mechanisms make the system resilient to evolving financial attack strategies.

The framework's modular design enables seamless integration into existing financial infrastructure, while its adaptive thresholding mechanism ensures continued effectiveness amid changing market conditions and attack sophistication levels. This study substantially reinforces the security of ML models in practical digital finance deployments, contributing to the broader goals of trustworthy AI in financial technology.

References

[1] 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," in *CEUR Workshop Proceedings*, vol. 2808, 2021.

[2] T. L. Melo, G. Ramos, and R. Meira, "Adversarial training for tabular data with attack propagation," in *KDD Workshop on Machine Learning in Finance*, arXiv:2307.15677, 2023.

[3] I. Fursov, M. Morozov, N. Kaploukhaya A, Kovtun R. Rivera, G. Gusev, B. Nikolenko, E. Burnaev, and R. Barannikov, "Adversarial attacks on deep models for financial transaction records," in *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery & Data Mining*, arXiv:2106.08361, 2021.

[4] P. Rani, "Making ML models in banking resilient using adversarial attacks," *Wipro AI & Analytics Insights*, 2022.

[5] A. Ojo, "Adversarial machine learning in finance: Developing resilient AI models to counter fraudster evasion," *Journal of Financial Cybersecurity and AI Robustness*, 2023.

[6] National Institute of Standards and Technology (NIST), "Adversarial machine learning: Taxonomy and terminology of attacks and mitigations," *NIST Artificial Intelligence Risk Management Framework (AI RMF)*, 2025.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)