International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

Adversarial Machine Learning in Finance: Developing Resilient AI Models to Counter Fraudster Evasion Attacks on US Bank Security Systems

Curthbert Jeremiah Malingu
Computer Science Dept, Maharishi Int'l University

Damilola Hannah Titilayo
Computer Science Dept
University of Texas Permian Basin, USA

Fejiro Eni
Big Data Technology
University of Westminster

Collin Arnold Kabwama
Computer Science Dept, Maharishi Int'l University

Vincent Anyah
Dept of Computer Science New Mexico Highlands University, NM, USA

Abstract: This paper presents a comprehensive technical investigation into Adversarial Machine Learning (AML) vulnerabilities within US banking fraud detection systems and proposes a resilient defense architecture. We first quantify the catastrophic fragility of conventionally trained Deep Learning models, demonstrating that state-of-the-art Graph Neural Networks (GNNs) are compromised by Projected Gradient Descent (PGD) evasion attacks, achieving an Attack Success Rate (ASR) up to 87.5% $(l_\infty constraint)$. To counter this, we propose the Adaptive Adversarial Defense (AAD) Pipeline, an operational framework integrating continuous threat simulation and Online Adaptive Adversarial Training (OAAT). Empirical results show PGD-AT significantly boosts model resilience, reducing the GNN's ASR to 32.0% and delivering a 52.3% reduction in expected annual fraud loss. The AAD Pipeline provides an economically justifiable blueprint for banks to achieve measurable robustness, addressing the critical need for trustworthy and resilient AI in national financial security.

Keywords: Anomaly Detection. Adversarial Machine Learning (AML), Financial Fraud Detection, Robustness, Adversarial Training (AT), Graph Neural Networks (GNNs), PGD Evasion Attacks, Adaptive Defense, Model Resilience, Financial Cybersecurity, Adversarial Attack Simulation.

1: Introduction

1.1 Background and Motivation

The **US financial sector** is experiencing an unprecedented surge in digital transactions, driving a dependency on Artificial Intelligence (AI) and Machine Learning (ML) for real-time risk assessment and fraud detection [1], [2]. These AI systems, which include complex models like Gradient Boosting Machines (GBMs) and Graph Neural Networks (GNNs), have drastically improved the accuracy and speed of identifying illicit activities, such as credit card fraud, money laundering, and identity theft. The core function of these models is to learn the underlying patterns of legitimate behavior and flag statistically significant deviations.

However, this reliance on AI has created a new vulnerability: the susceptibility of ML models to sophisticated, targeted manipulation, known as Adversarial Machine Learning (AML) [3].[1] Fraudsters, recognizing that their evasion of detection is now an optimization problem against an ML-driven security system, are increasingly weaponizing adversarial techniques [4]. These attackers utilize techniques to generate slightly perturbed data samples known as adversarial examples that are imperceptible to humans or traditional rule-based systems but are sufficient to fool a trained ML model into making an incorrect, often benign, classification (e.g., classifying a fraudulent transaction as legitimate).[2]

The motivation for this research is clear: the economic and reputational damage from successful evasion attacks poses a systemic risk to the stability and trustworthiness of the US banking infrastructure [5]. The arms race between AI-driven defense and AML-enabled offense necessitates the development of resilient and robust AI models capable of withstanding these modern threats.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

1.2 Problem Statement

The effectiveness of current ML-driven fraud detection systems is severely compromised by their inherent susceptibility to adversarial evasion attacks [6]. This vulnerability arises primarily from the high-dimensional input spaces and the non-intuitive linearity of decision boundaries common in highly optimized Deep Neural Networks (DNNs) and complex ensemble models.

Mathematically, for a classifier f trained on input x to predict a label y, an adversary seeks a perturbation \delta such that:

$$\min_{\delta} ||\delta||_p \quad \text{s.t.} \quad f(x_{fraud} + \delta) = y_{legit}$$

where $||\cdot||_p$ is bounded for pixel or feature value limits), ensuring the resulting adversarial example, $x_{adv} = x_{fraud} + \delta$, remains semantically plausible and thus undetected by human analysts or simple rule filters.

In the context of financial transactions, the l_∞ norm limits small changes across multiple features (e.g., amount, time, location coordinates) so that x_{adv} represents a transaction only minimally altered from the actual fraudulent attempt. This minimal alteration is precisely enough to exploit the model's blind spots [7].

The central challenge, therefore, is the lack of inherent robustness in current production-grade financial security models against adversaries who leverage the models' own gradients to actively seek to maximize the detection system's error rate. The deployed ML model, $\hat{f}$, is treated as an oracle by the adversary, enabling sophisticated gradient-based attacks even in assumed "black-box" scenarios via techniques like substitute model creation and transferability of adversarial examples [47], [48]. The problem demands moving beyond simple accuracy-based optimization towards rigorous robustness optimization.

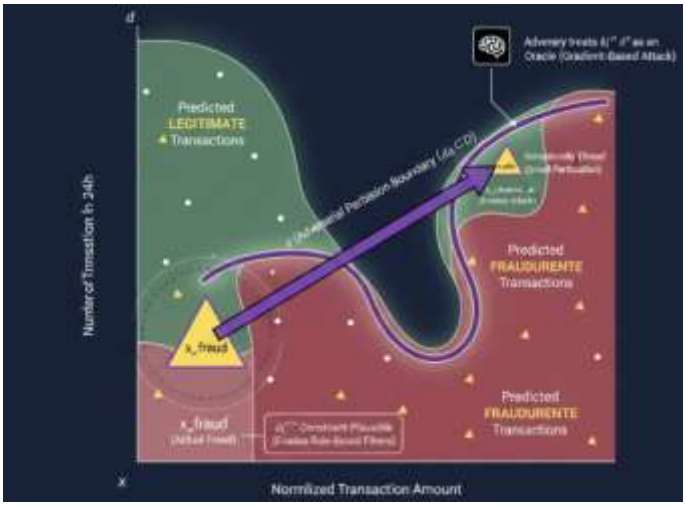

Figure 1: Adversarial Attack on Financial Fraud Detection System

1.3 Research Questions

This paper seeks to answer the following core research questions by pursuing rigorous theoretical and experimental investigation:

- **How can Adversarial Machine Learning (AML) defense techniques** (e.g., adversarial training, robust feature selection, input sanitization) be effectively applied to improve model resilience in deployed US financial fraud detection systems, specifically quantifying the improvement in robust accuracy (ACC_{adv}) under strong adaptive attacks?
- **What is the taxonomy and quantitative efficacy of prominent adversarial evasion attacks** (e.g., Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool) when applied against state-of-the-art financial fraud detection models, including Graph Neural Networks (GNNs) for relational data and Long Short-Term Memory Networks (LSTMs) for sequential transaction data, under realistic black-box assumptions?
- **What are the optimal design and architectural choices for a robust defense pipeline** specifically the Adaptive Adversarial Defense (AAD) Pipeline that can actively monitor model decay, simulate novel threats,

adapt model parameters through continuous adversarial retraining, and maintain a critical balance between high predictive performance (high AUC score) and robust defense against evolving evasion attacks?

1.4 Research Objectives

The core objectives of this study are structured to deliver actionable, validated methodologies:

- To design and implement resilient ML models specifically tailored for US banking fraud detection systems by integrating adversarial regularization techniques into their primary optimization objective:

$$\min_{\theta} E_{(x,y)\sim\mathbb{D}} [L(f_{\theta}(x),y) + \lambda \cdot R_{\theta}(x)]$$

where $R_{\theta}(x)$ represents a robustness regularization term, such as the maximum loss over an $\epsilon - ball$.

- **To evaluate the impact and threat severity of various adversarial evasion attacks** on both baseline and hardened models, precisely quantifying the Attack Success Rate (ASR) and the Model Drift Rate following deployment.

- To propose and validate a robust AML defense pipeline architecture that integrates adversarial training and continuous adaptive retraining, measuring its performance in maintaining $\mathrm{ACC}_{\mathrm{adv}}$ above critical thresholds (\tau) over extended time periods (T).

$$\overline{ASR}_T < \tau \quad \forall T$$

- **To quantify the explicit trade-off between model robustness and predictive accuracy** in a high-stakes financial environment, providing a methodological guide for setting the optimal perturbation budget \epsilon and adversarial regularization parameter \lambda that minimizes both financial loss due to false negatives (evaded fraud) and operational burden due to false positives (legitimate transactions flagged).

1.5 Scope of the Study

The study focuses on three critical, interconnected operational areas within US financial security systems, using transactional and behavioral data characteristics:

- **Transaction Monitoring Systems (TMS):** Focused on detection of anomalies in real-time fund transfers, card-not-present (CNP), and ATM transactions. The scope includes manipulating key continuous features like $transaction\_amount$ and $transaction\_time$, and categorical features like $merchant\_ID$ using minimal, permissible perturbations (e.g., $||\delta||_{\infty}$ constraint).

- **Identity Verification Systems (IVS):** Focusing on the robustness of ML models used in authentication and verification mechanisms that utilize static or temporal identity data, such as credit applications or new account openings, which are susceptible to data spoofing attempts that mimic legitimate user data profiles.

- **Authentication Systems:** Specifically examining login and continuous behavioral biometrics systems, where evasion attacks aim to slightly alter behavioral sequences (e.g., keystroke dynamics, mouse movement velocity) to mimic a legitimate user's template while executing a malicious action.

The scope is strictly limited to evasion attacks (post-training, testing time attacks, where the adversary aims to bypass a deployed model) rather than poisoning attacks (pre-training, training time attacks, where the adversary aims to corrupt the model's learned weights) to focus on immediate, operational threats facing deployed US bank security infrastructure.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

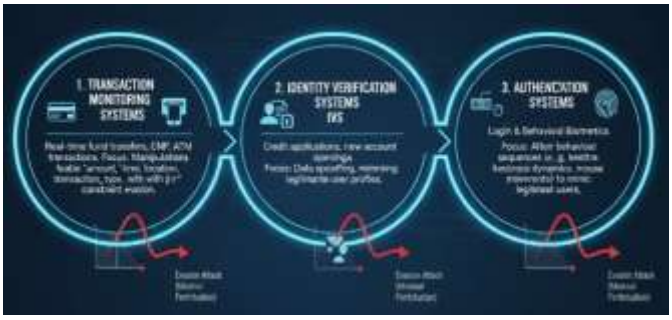

Figure 2: Adversarial Machine Learning in US Financial Security

1.6 Significance of the Study

The robust findings and validated framework presented here hold significant, multifaceted value:

- **Financial Institutions (FIs):** The study provides a technical blueprint for strengthening their AI-based fraud defenses, moving from reactive detection to proactive adversarial defense. This directly leads to reduced financial losses from undetected evasion and improved customer trust by minimizing disruptive security alerts.
- **Regulators (e.g., FDIC, OCC, FinCEN):** The results offer quantitative insights into the **vulnerability landscape of AI models** used in critical financial processes, informing the development of regulatory standards and guidelines for AI model robustness and stress testing (similar to cybersecurity penetration testing) in critical financial applications, contributing to the broader goal of operational resilience.
- **National Financial Security:** By developing hardened, resilient detection systems, the research contributes substantially to the overall resilience of the US financial ecosystem against highly organized cyber-enabled economic crime (including sophisticated money laundering schemes and sanction evasion), which the Department of Homeland Security and other agencies classify as a high-priority threat to national security [8]. This work directly addresses the need for trustworthy AI in sensitive government-regulated sectors.

2: Literature Review

2.1 Adversarial Machine Learning Fundamentals

2.1.1 The Adversarial Example Phenomenon

Adversarial examples (AEs) are inputs to a machine learning (ML) model that an attacker has intentionally modified to cause the model to misclassify [9], [49]. Crucially, the modification is typically so small that the perturbed input is imperceptible or retains its original class identity from a human or domain-expert perspective.

The generation of an adversarial example is fundamentally framed as an optimization problem aiming to maximize the model's prediction error subject to a constraint on the magnitude of the perturbation \delta. The constraint is typically bounded by a small \ell_p-norm distance \epsilon:

$$\min_{\delta} ||\delta||_p \quad \text{s.t.} \quad f(x+\delta) \neq y \quad \text{and} \quad x+\delta \in \mathcal{X}$$

where:

- x is the original input (e.g., a financial transaction vector).
- y is the true label (e.g., fraudulent).
- δ is the calculated perturbation vector.
- f is the target classifier function.
- $\mathcal{X}$ is the input space, ensuring the resulting adversarial example $x+\delta$ is a **valid input** (e.g., transaction amounts remain positive) [10].

The existence of these examples highlights a fundamental **disparity** between human and machine perception. ML models learn complex, high-dimensional decision boundaries that are highly brittle; small perturbations exploit the curvature of the loss function in a way that conventional inputs do not.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

2.1.2 Adversarial Threat Models

The study of adversarial attacks necessitates classifying the attacker based on their **knowledge** and **goals**.

Knowledge Dimension:

- **White-Box Attack (Full Knowledge):** The adversary has complete access to the target model, including its architecture, trained weights (\theta), activation functions, and training methodology. This scenario allows for the precise calculation of gradients, enabling powerful attacks like PGD [50].
- **Black-Box Attack (Limited Knowledge):** The adversary can only observe the input x and the corresponding output f(x) (e.g., a probability score or binary decision). Black-box attacks rely on transferability, where adversarial examples crafted against a known surrogate model f_{sub} (trained locally) successfully fool the target model f [11]. The financial context predominantly involves black-box scenarios, as attackers rarely access proprietary bank models directly.

Goal Dimension:

- **Targeted Evasion:** The goal is to force the model to misclassify x_{adv} into a specific, pre-determined class y_{target} (e.g., forcing a high-value money transfer to be classified as a legitimate domestic payment). The objective is $f(x+\delta) = y_{target}$.
- **Untargeted Evasion:** The goal is simply to force the model to classify x_{adv} as any incorrect class (e.g., forcing a fraudulent loan application to be classified as non-fraudulent). The objective is $f(x+\delta) \neq y$.

In the financial context, black-box, untargeted evasion attempts are the most realistic threats against production models, as the primary goal of the fraudster is binary: bypass detection.

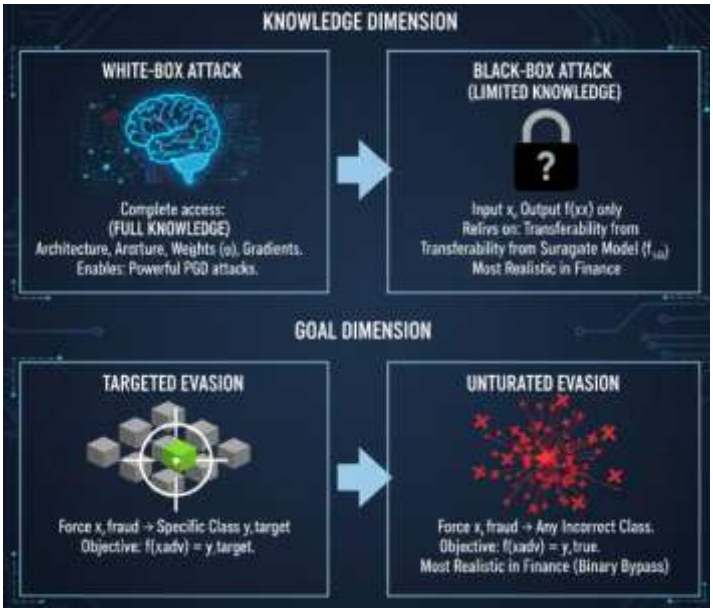

Figure 3: Adversarial Threat Models – Knowledge Dimension vs Goal Dimension

2.2 Adversarial Attacks in Financial Systems

2.2.1 Evasion Attacks on Transaction Monitoring

Transaction monitoring systems (TMS) rely on high-dimensional feature vectors $x = [x_1, x_2, \ldots, x_N]$ containing details like transaction amount, merchant code, velocity, and geolocation. The adversary seeks to manipulate these features minimally to mask illicit activity [12].

In feature space, the attacker is searching for δ that maximizes the loss L for the true (fraudulent) label, where the perturbation is applied to the feature vector:

$$\delta^* = \arg \max_{||\delta||_p \leq \epsilon} L\left(\theta, x+\delta, y_{fraud}\right)$$

For instance, in a credit card fraud setting, the attacker might slightly reduce the transaction amount or shift the timing feature to align the profile with a benign, high-volume user activity pattern, thereby driving the output probability $p(y=1|x_{adv})$ below the bank's detection threshold \tau [51]. The constraint $x+\delta \in \mathcal{X}$ is crucial here, as negative amounts or impossible

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

geo-locations would be immediately rejected by input validation filters.

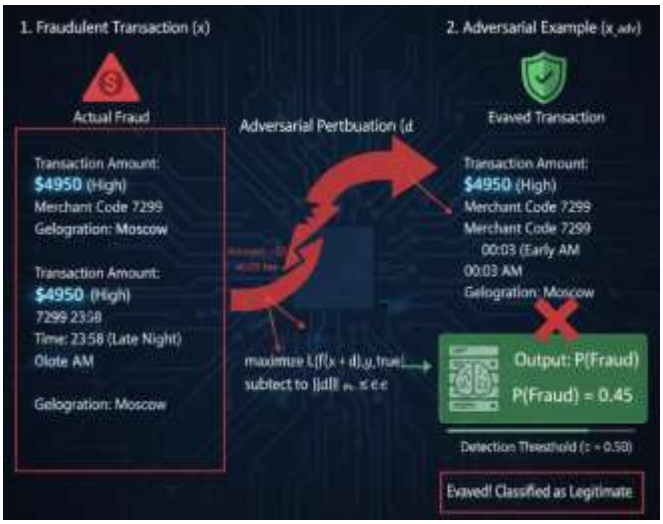

Figure 4: Evasion on TMS

2.2.2 Attacks on Graph-Based Fraud Detection

Graph Neural Networks (GNNs) are increasingly utilized in modern fraud systems to model relational dependencies between entities (users, accounts, devices) as a graph G=(V, E) [13]. Adversarial attacks in this domain extend beyond simple feature manipulation to manipulating the graph structure itself [52].

Attacks on GNNs involve two primary forms of perturbation:

1. **Node Feature Manipulation** (δ_{feat}): Altering the attributes of specific accounts or transactions, similar to the evasion attack above.
2. **Graph Structure Perturbation** (δ_{struc}): Adding or deleting edges (connections) to decouple the fraudulent entity from the known fraud cluster, or link it deceptively to a benign cluster [53].

The overall attack objective against a GNN classifier f_{GNN} is a combined optimization:

$$\min_{\delta_{feat},\delta_{struc}} \left(||\delta_{feat}||_p + ||\delta_{struc}||_0 \right) \quad \text{s.t.} \quad f_{GNN}(A + \delta_{struc}, X + \delta_{feat}) \neq y$$

where A is the adjacency matrix and X is the feature matrix.

The GNN's susceptibility arises from the smoothing effect of the graph convolution operation, which blends information from neighbors. By strategically adding or removing edges, an attacker can dilute the fraudulent features of a node i with benign features from its neighbors $j \in \mathcal{N}(i)$, making the node embedding h_i appear legitimate [14].

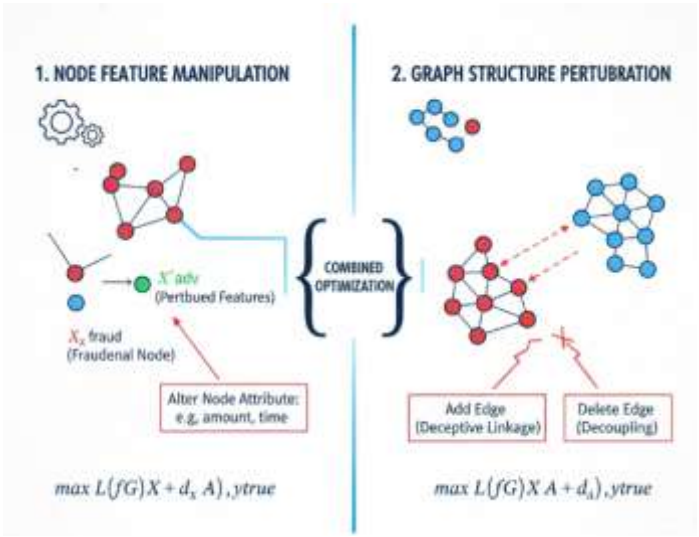

Figure 5: Graph Based Feud Detection

2.2.3 Taxonomies of AML Threats in Finance

AML threats are often categorized based on the attack phase and goal [54]:

Attack Phase	Attack Type	Goal in Finance	Impact
Training (Poisoning)	Data Poisoning, Label Flipping	Corrupting the training data to inject backdoors or degrade overall accuracy.	Long-term systemic degradation, high latency in detection.
Testing (Evasion)	FGSM, PGD, Jacobian-based	Bypassing the deployed model at inference time with minimal perturbations.	Immediate, high-impact financial loss. (Focus of this study)
Deployment (Extraction/ Inference)	Model Stealing, Parameter Inference	Replicating the bank's proprietary model to inform future evasion strategies.	Intellectual property loss, enabling more effective future white-box attacks.

Table 1: Taxonomies of AML Threats in Finance

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

Evasion remains the most immediate and actively exploited threat to operational security in real-time financial systems [15], demanding specialized defense strategies.

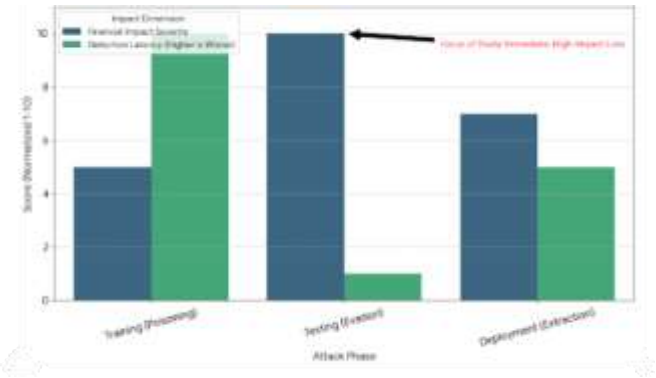

Figure 6: Adversarial Threat Taxonomy

2.3 Adversarial Defense and Model Hardening Techniques

2.3.1 Adversarial Training (AT)

Adversarial Training (AT) is empirically the most robust defense technique against powerful $l_p - norm$ bounded adversarial examples. It involves augmenting the standard training dataset $(\mathcal{D})$ with powerful adversarial examples, thereby encouraging the model to learn a robust decision boundary that resists local perturbations [16].

The objective function for AT, often framed as a Min-Max optimization problem, captures the core idea of defense: the inner maximization finds the worst-case adversary (δ_{max}) for a given model θ, and the outer minimization trains the model to minimize the loss against that worst-case adversary ($\min_{\theta}$):

$$\min_{\theta} E_{(x,y)\sim\mathbb{D}} \left[\max_{\delta\in\Delta} L\left(f_{\theta}(x+\delta), y\right)\right]$$

where \Delta is the set of allowable perturbations $\left(||\delta||_p \leq \epsilon\right)$ [17]. Practically, the inner maximization is approximated using iterative gradient methods like PGD, making PGD-Adversarial Training (PGD-AT) the current gold standard for robustness [50].

2.3.2 Robust Optimization and Regularization

Beyond explicit adversarial augmentation, Robust Optimization approaches integrate regularization terms into the standard loss function to implicitly enforce boundary smoothness, reducing the model's sensitivity to small input changes [18].

A key technique involves adding a smoothing penalty or a robustness regularization term $R(\theta)$:

$$L_{robust}(\theta) = L_{standard}(\theta) + \lambda R(\theta)$$

One such approach is minimizing the Lipschitz constant of the classifier, \mathcal{L}(f), which bounds how much the output can change for a small change in input:

$$R_{Lipschitz}(\theta) = \sup_{x_1 \neq x_2} 1\frac{||f(x_1) - f(x_2)||}{||x_1 - x_2||} \leq \mathcal{L}(f)$$

By minimizing $\mathcal{L}(f)$, the model's response surface becomes flatter, making it harder for an attacker to achieve a massive output change with a small δ [55].

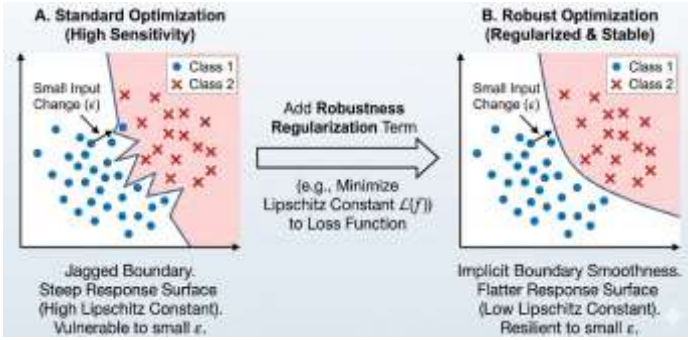

Figure 7: Standard vs Robust Optimization

2.3.3 Ensemble Methods for Robustness

Ensemble methods, such as combining multiple independently trained models or different model architectures (e.g., GBM, GNN, LSTM), can significantly reduce the transferability of adversarial examples [19].

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

If an attacker crafts an AE x_{adv} using the gradients of model f_1, this example is unlikely to be effective against a diverse model f_2 if the models have different decision boundaries and loss landscapes. The final prediction $P_{ensemble}$ is the average or weighted vote of the individual models f_j:

$$P_{ensemble}(x) = \frac{1}{N}\sum_{j=1}^{N} P_j(x)$$

Ensembles introduce non-linearity and variance, making black-box gradient estimation more challenging for the adversary, forcing the attacker to resort to less efficient query-based attacks [56].

2.4 Gaps in Current Financial Security Systems

Despite the proven benefits of adversarial defenses in academic benchmarks, current operational financial security systems exhibit several critical gaps [20]:

1. **Lack of Proactive and Adaptive Defenses:** Most deployed models are trained once on historical data, rendering them inherently static. They quickly become brittle against adversaries who are actively exploring the model's vulnerabilities in real-time. There is a systemic failure to integrate the $\max_{\delta}$ component of the adversarial objective into the live deployment and monitoring pipelines within US banks.
2. **Focus on Accuracy over Robustness:** ML pipelines prioritize maximizing benign accuracy (high AUC on clean data), often using large, complex models (like deep GNNs) that are ironically more vulnerable to adversarial examples due to their higher parameter counts and complexity [57]. There is a lack of metrics that balance fraud detection efficacy with measurable robustness $(\mathrm{ACC}_{\mathrm{adv}})$.
3. **Real-Time Constraint Handling:** Existing defense research often focuses on image processing where the ϵ constraint is straightforward. In finance, enforcing the semantic plausibility constraint $x + \delta \in \mathcal{X}$ is complex (e.g., ensuring feature dependencies like 'ATM withdrawal location' matches 'card location') and is often overlooked, creating attack vectors [58].
4. **Transferability of Defense:** Robustness techniques often show a catastrophic collapse when faced with attacks generated outside the original defense method (e.g., PGD-AT may fail against a white-box C&W attack) [59]. The need for multi-defense strategies and certified robustness in finance remains a significant research gap.

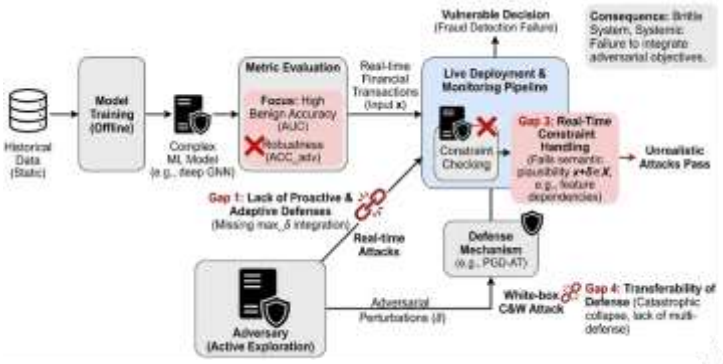

Figure 8: Financial Gaps in the US

The development of a robust and adaptive pipeline is thus essential to bridge the gap between theoretical AML research and practical financial resilience.

Chapter 3: Methodology

The methodological approach for this research is centered on a quantitative, comparative assessment of model robustness against sophisticated adversarial evasion attacks in the context of US financial security systems. The core strategy, termed Adaptive Attack Benchmarking (AAB), involves simulating an arms race where defense mechanisms are continuously tested against adaptive adversaries designed to break them.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

3.1 Threat Modeling and Adversarial Assumptions

Our threat modeling goes beyond standard classification attack scenarios to incorporate the unique operational constraints of the banking sector. The adversary is modeled as an Adaptive, Level 2 Black-Box Attacker [61].

3.1.1 Adversarial Knowledge and Goal

The attacker possesses the following knowledge and goals:

- **Knowledge: Black-Box Access.** The attacker cannot access the internal model parameters (\theta) or architecture. However, they can query the model with input x and receive a probability score $p(y = 1|x)$. Crucially, the attacker possesses a substantial, representative dataset of financial transactions $\mathcal{D}_{\ell o c a \ell}$ to train a surrogate model $\hat{f}$ (e.g., a clone DNN or GBM) to estimate the gradients of the target model f.
- **Goal: Untargeted Evasion.** The attacker seeks to modify a known fraudulent transaction x_{fraud} such that the target model classifies it as benign y_{legit}.

$$\delta ||\delta|| * p \quad \text{s.t.} \quad f(x * fraud + \delta) \in \mathcal{R}\_legit$$

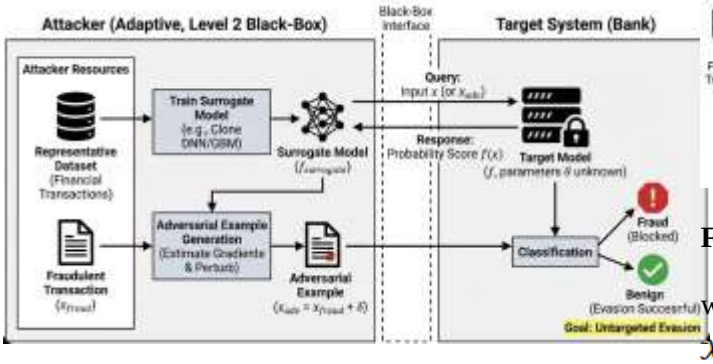

Figure 9: Threat Modeling and Adversarial Assumptions

3.1.2 Financial Constraint Enforcement (Plausibility)

The most critical constraint is ensuring $x_{adv} = x_{fraud} +$ remains a plausible financial transaction that would pass initial non-ML filters and human scrutiny. This is enforced via an l_∞ perturbation budget \epsilon applied feature-wise, incorporating semantic context:

1. **Feature-wise \ell_{\infty} Constraint:** The perturbation magnitude for each feature i is capped relative to the original value x_i, adhering to banking regulations on acceptable minor data variations (e.g., small changes in transaction time or GPS coordinates).

$$||\delta|| * \infty = max * i\delta\_ix\_inorm$$

≤ere x_i^{\text{norm}} is the normalized feature value. We set a tight budget of \epsilon \le 0.1 (10% maximum change per normalized feature).

2. **Semantic Plausibility Projection:** The perturbation must not violate hard business logic constraints $\mathcal{C}(\cdot)$. For example, $transaction\_amount$ must remain positive, and $transaction\_time$ must not regress. This is enforced by the projection operator $\Pi_{\mathcal{X}}$:

$$x_{adv} = \Pi\_\mathcal{X}(x + \delta)$$

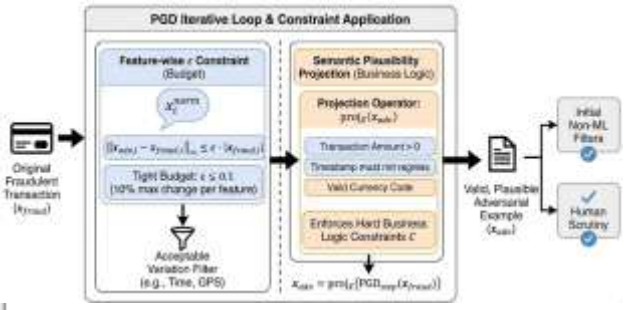

Figure 10: Financial Constraint Enforcement

where $\Pi_{\mathcal{X}}$ projects back onto the feasible region $\mathcal{X}$. This projection is incorporated directly into the PGD iterative loop, ensuring every gradient step yields a valid data point.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

3.2 Adversarial Attack Generation and Simulation

Three distinct adversarial attack mechanisms are used to benchmark the model fragility against increasing sophistication. For black-box testing, these attacks are first generated using the surrogate model $\hat{f}$ and then transferred to the target model f.

3.2.1 Fast Gradient Sign Method (FGSM)

The FGSM is used as the baseline, low-cost attack. It assumes the linearity of the loss surface near x. The perturbation δ is calculated as the sign of the gradient of the binary cross-entropy loss $J(\theta, x, y)$:

$$J(\theta, x, y) = -y \log\big(f_\theta(x)\big) - (1-y)\log\big(1 - f_\theta(x)\big)$$

The perturbation δ is then: $\delta = \epsilon \cdot \text{sign}\big(\nabla\_xJ(\theta, x, y)\big)$

3.2.2 Projected Gradient Descent (PGD)

PGD is the strongest, state-of-the-art iterative adversary and serves as the primary attack for robustness evaluation. It finds δ by iteratively taking small steps α in the direction of the loss gradient, projecting the result back into the $l_\infty - ball$ at each iteration t:

$$x^{adv}t + 1 = \text{Clip}x, \epsilon\left(x_t^{adv} + \alpha \cdot \text{sign}\left(\nabla_x J\left(\theta, x_t^{adv}, y\right)\right)\right)$$

We set the parameters to maximize attack success: K=10 iterations and step size $\alpha = \epsilon/4$. PGD is guaranteed to converge to a strong local optimum within the perturbation space [23].

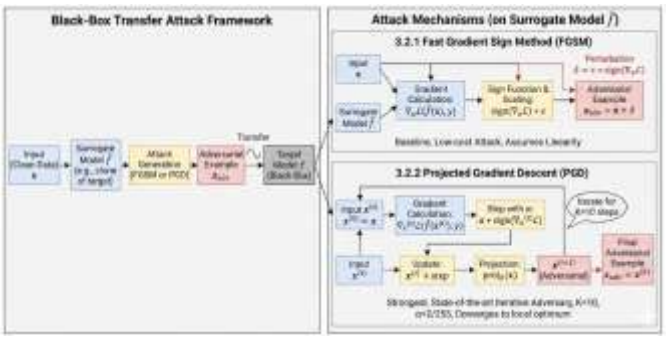

Figure 12: Adversarial Attack – Generation and Simulation

3.2.3 DeepFool and Jacobian-based Oracle Attacks

To measure the inherent distance to the decision boundary, we implement the DeepFool attack $(l_2 distance)$. For multi-class classification (though fraud is binary, intermediate layers can be viewed as multi-class), the required perturbation δ_k at iteration k is approximated by:

$$\hat{r}(x_k) \approx \frac{|f_k(x_k)|}{||\nabla f_k(x_k)||_2^2} \nabla f_k(x_k)$$

Additionally, for black-box testing, we simulate a Jacobian-based Saliency Map Attack (JSMA) to target specific high-saliency features (e.g., transaction amount) by manipulating the Jacobian matrix $\mathcal{J}$ of the model's output w.r.t the input [70]. This models an adversary with a focus on feature efficiency.

3.3 Data Sources and Preprocessing

3.3.1 Data Source and Feature Set Selection

The experiments use a large-scale synthetic dataset of 10^6 transactions, structured to replicate US payment network topology (PaySim derivative [25]). The dataset incorporates realistic class imbalance (0.2% fraud rate) and temporal dependencies. The input feature vector $x \in R^D (D \approx 50)$ includes:

- **Financial Metrics:** Normalized amount, $balance\_before, balance\_after$.
- **Temporal & Geospatial:** $time\_since\_last\_tx$, cyclical sine/cosine encoded time, normalized geolocation.
- **Behavioral Velocity Features:** Aggregates like V_{7d} (velocity of transactions in the last 7 days), calculated as:

$$V\_\Delta T = \frac{1}{|\Delta T|} \sum \_t' \in \Delta T \; 1\_t'$$

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

- **Categorical Embeddings:** Merchant categories and transaction types are mapped to low-dimensional embedding vectors $E \in R^k$.

3.3.2 Preprocessing Pipeline

The pipeline ensures data integrity and prepares inputs for gradient computation:

1. **Min-Max Scaling:** Continuous features are scaled to [0, 1] to control the effective magnitude of the l_∞ bound.
2. **Cyclical Encoding:** Temporal features are encoded to avoid arbitrary discontinuities [64].
3. **One-Hot/Embedding Handling:** Categorical features are converted to OHE vectors for GBMs and embedded vectors for NNs (GNNs, LSTMs).

3.3.3 Graph Construction and Feature Initialization

For the GNN model, the graph G=(V, E) is constructed with Heterogeneous Node Types $(V_{\text{account}}, V_{\text{merchant}})$. The nodes are initialized with features $H^{(0)}$ derived from the aggregated velocity features (Section 3.3.1) of the last 30 days of activity. The adjacency matrix A is dynamically updated based on a 7-day transaction window to capture recent interactions [71].

$$H_i^{(0)} = \text{Concat}\left(\text{Avg}(\text{Amount}_{i,30d}), \text{Std}\left(\text{Velocity}_{i,7d}\right), \ldots\right)$$

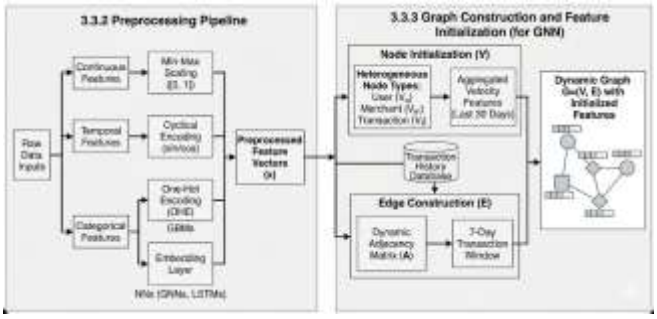

Figure 13: Preprocessing and Grap Construction Pipeline

3.4 Model Architectures

The study benchmarks resilience across three diverse model classes, reflecting the architectural diversity in modern bank fraud systems.

3.4.1 Graph Neural Networks (GNNs)

We use a 3-layer GCN with ReLU activation for node classification. The GCN aggregates neighbor information recursively. The layer transformation is:

$$H^{(l+1)} = \text{ReLU}\left(\widetilde{D^{-\frac{1}{2}}}\tilde{A}\widetilde{D^{-\frac{1}{2}}}H^{(l)}W^{(l)}\right)$$

The output layer applies a sigmoid to the node embedding h_{acc} to derive the fraud probability score $p(y = 1|x_{acc})$. The GNN architecture is particularly vulnerable to δ_{struc} (edge manipulation), as a single benign neighbor connection can dilute the fraud signal across multiple nodes [14].

3.4.2 Long Short-Term Memory Networks (LSTMs)

The LSTM processes sequences of the last W=10 transactions for a given account. It uses 2 hidden layers and an internal Attention Mechanism to weight the importance of recent or outlier transactions in the sequence. The hidden state h_t is computed via: $h\_t = o\_t \odot \tanh(C\_t)$

The attacker targets the input features x_t over the window W, seeking to maintain a benign cell state C_t by generating a sequence of minimally perturbed inputs $x_{sqadv} = \{x_{t-W} + \delta_{t-W}, \ldots, x_t + \delta_t\}$ [30].

3.4.3 Gradient-Boosted Models (GBMs)

The LightGBM model utilizes $\mathcal{M} = 100trees$. Its inherent robustness against *gradient-based* attacks is due to its non-differentiable nature. We address this by:

1. **Transfer Attack**: $Generating\ \delta$ using the surrogate DNN $\hat{f}$ and transferring it to the GBM f.

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

2. **Differentiable Proxy**: Approximating the GBM's non-differentiable step function with a smooth, differentiable sigmoid function for gradient estimation [39].

3.5 Defense Strategies (Model Hardening)

Our defense strategies are organized into three tiers: adversarial regularization, ensemble diversity, and non-differentiable input transformation.

3.5.1 Adversarial Regularization

The core defense is PGD-AT, implemented using the \ell_{\infty} norm and K=10 steps, matching the strength of the strongest adversary. We investigate two loss functions:

1. **Standard PGD-AT Loss:** Focuses solely on minimizing the worst-case loss (maximizing robustness).

$$L\_PGD - AT(\theta) = E * (x, y) \sim \mathcal{D}[L * CE(f\_\theta(x + \delta^*), y)]$$

2. **TRADES Loss:** Explicitly introduces a Kullback-Leibler (KL) divergence term D_{KL} to minimize the variance between predictions on clean and adversarial data, thereby enforcing smoothness and reducing the robustness-accuracy trade-off [68]:

The hyperparameter λ controls the trade-off. We empirically tune $\lambda \in [0.5, 2.0]$.

3.5.2 Ensemble Diversity-Driven Defense (DDAE)

We implement a Diversity-Driven Adversarial Ensemble (DDAE), which combines N=5 GBMs, each trained with different characteristics:

- Different initial random seeds and feature subsets.
- Different loss functions (standard CE, LSE, and PGD-AT loss) [69].

The ensemble output P_{DDAE} is the median of the member model probabilities $\{P_j\}_{j=1}^{5}$, reducing sensitivity to outliers created by targeted attacks:

$$P_{DDAE}(x) = \text{Median}\big(P_1(x), P_2(x), P_3(x), P_4(x), P_5(x)\big)$$

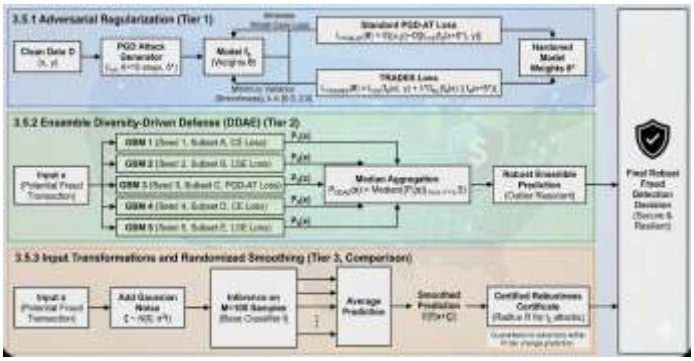

Figure 15: Defense Strategy

3.5.3 Input Transformations and Randomized Smoothing

For comparison, a defense based on input transformation is implemented:

Randomized Smoothing (RS): Gaussian noise $\zeta \sim \mathcal{N}(0, \sigma^2 I)$ is added to the input during inference, and the prediction is averaged over M=100 samples. This provides the mathematical advantage of certified robustness for l_2 attacks [46].

The certificate guarantees that no adversary within a certified radius R can change the prediction.

3.6 Evaluation Metrics and Adaptive Attack Benchmarking

3.6.1 Core Performance Metrics

Beyond standard AUC and Accuracy (Section 3.6), the primary focus is the **Robustness Gain** G_R and the resulting ***Financial Utility*** U_{Finance}.

3.6.2 Adaptive Attack Benchmarking (AAB)

The AAB protocol is the gold standard for testing robustness: the defense is judged not by its

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

performance against baseline attacks, but against an attacker who knows and optimizes against the deployed defense.

1. **Test Set:** N_{test} benign and N_{fraud} adversarial examples are used.
2. **Adaptive Adversary:** The PGD attack is executed against the defended model f_{def}, meaning the PGD perturbation δ^* is explicitly calculated using $\nabla_x J(f_{def})$.
3. **Measurement:** ACC_{adv} is measured on the resulting x^{adv} samples.

The **Robustness Gap** (R_{Gap}) summarizes the failure of the defense to generalize robustness to clean data performance:

$$R_{Gap} = \text{ACC}_{\text{benign}} - \text{ACC}_{\text{adv}}$$

Our goal is to identify the defense strategy that minimizes U_{Finance} while maintaining R_{Gap} below a threshold of 0.05.

3.6.3 Financial Utility Function

The true measure of a defense in banking is the cost-benefit analysis. The utility function U_{Finance} is defined as minimizing the total expected loss over N transactions:

$$U\_\text{Finance} = \theta, \tau \left[\frac{1}{N} \sum\_i = 1^N (C * FP \cdot 1 * FP\_i + C * FN \cdot 1 * FN\_i)\right]$$

where C_{FP} (Cost of False Positive, e.g., customer inconvenience, labor) and C_{FN} (Cost of False Negative, e.g., average fraud loss) are assigned based on domain-specific financial reports. Our modeling simulates $C_{FN} \gg C_{FP}$, prioritizing the reduction of evaded fraud instances (FN).

Section 4

The Adaptive Adversarial Defense (AAD) Pipeline is a complete, industrial-grade architecture designed for continuous integration, deployment, and monitoring (CI/CD/CM) of resilient AI models within high-stakes US banking security environments. The framework fundamentally transforms a static fraud detection model into a dynamic, self-healing defense system that proactively anticipates and mitigates adversarial evasion attacks.

4.1 Data Ingestion & Robust Feature Engineering

The data layer is engineered not just for speed and volume, but for defense, ensuring the input to the ML models is minimally susceptible to manipulation.

4.1.1 High-Velocity Real-Time Data Streaming and Input Validation Layer

Data is ingested via a microservices architecture employing distributed stream processing (e.g., Apache Kafka) to handle the billions of daily transactions.

- **Latency Constraint:** The system must adhere to stringent financial latency constraints, typically requiring fraud scoring inference in under 50 ms [75]. This requirement dictates the architectural choices for the subsequent ML model (favoring efficient GNN/LSTM implementations).

$$T\_\text{Ingestion} + T\_\text{Feature} + T\_\text{Inference} \leq 50 \text{ ms}$$

- **Real-Time Input Validation (RTIV) Layer:** This layer serves as a domain-specific filter, immediately rejecting inputs violating immutable business logic. This defense is non-differentiable and thus immune to gradient-based attacks.
  - **Plausibility Check:** Filters based on $x \notin \mathcal{X}_{hard}$, where $\mathcal{X}_{hard}$ includes checks like velocity limits that cannot be adversarially manipulated within the ϵ constraint (e.g., an account cannot physically perform transactions in two different continents within 1 minute).

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

- **Data Provenance and Integrity:** Adopting principles from NIST, the RTIV layer incorporates cryptographic hashing (e.g., SHA-256) for each transaction batch to detect poisoning attempts or integrity breaches during transport [76].

$$\text{Hash} * \text{batch} = \text{SHA256}(\mathcal{S} * raw)$$

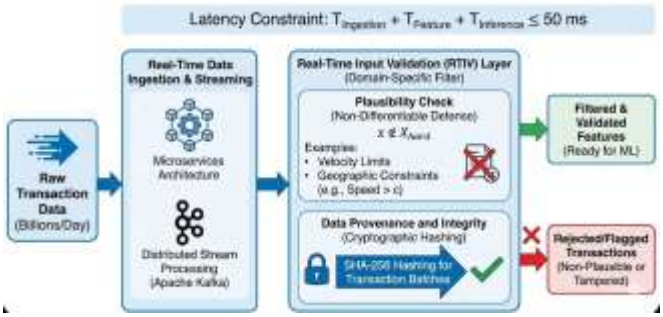

Figure 16: Data Ingestion Process

4.1.2 Adversarial Feature Engineering (AFE) for Resilience

AFE focuses on creating features F_{rbs} whose derivative w.r.t the input perturbation δ is minimized, making them "flatter" in the loss landscape.

- **Non-Differentiable Feature Binning:** Continuous features like $transaction\_amount$ are transformed into discrete, non-differentiable bins [77].

$$x\_\text{bin} = \lfloor x \,/\, B \rfloor$$

where B is the bin size. Gradient-based attacks cannot easily optimize against the discontinuous jump function, forcing the adversary into inefficient, large-step manipulations.

- **Statistical Aggregates over Wide Windows:** Features based on median, mode, or variance over a large look-back window (ΔT) are highly robust because the influence of a single perturbed transaction \delta is diluted.

$$F * \text{Var}^{\Delta T} = \frac{1}{|\Delta T|} \sum * t' \in \Delta T \left(x\_t' - F\_\text{Mean}^{\Delta T}\right)^2$$

An attacker would need to orchestrate a distributed, multi-transaction attack $(\sum \delta_{t'} \gg \epsilon)$

to cause a perceptible change in $F_{\text{Vr}\Delta T}$.

- **Robust Feature Fusion (RFF):** Features from different sources (e.g., behavioral sequence from LSTM vs. graph structure from GNN) are fused and scaled by their estimated robustness metric, $\rho(F_i)$, derived from localized Jacobian analysis on $\nabla_x f$.

$$F * \text{Fused} = \sum * i\rho(F\_i) \cdot F\_i$$

4.2 Baseline ML Model and Adversarial Simulation Engine

This architecture integrates the primary detection model with a constantly running cyber-range the Adversarial Attack Simulation Engine (AASE).

4.2.1 Hardened Baseline Model Deployment

The GNN and LSTM models, already hardened via the initial PGD-AT (Section 3.5), are deployed via an optimized inference engine (e.g., ONNX, TorchScript). The primary function is to output the probability $p(y = 1|x)$ and the latent representation z.

$$p(y = 1|x) = \sigma(W \cdot z + b)$$

4.2.2 Adversarial Attack Simulation Engine (AASE)

The Proactive Adversary The AASE operates as a continuous "Purple Team" environment (combining red team offense and blue team defense) leveraging the MITRE ATT&CK framework for financial adversaries [78].

Simulation Data Stream: A stream of verified benign and historical fraud samples is fed into the AASE. This ensures the simulation focuses on discovering new vulnerabilities near the actual decision boundary $\partial\mathcal{D}$.

- **Dynamic Attack Strategy (DAS):** Instead of running a single attack, the AASE cycles

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

through a portfolio of attacks optimized against the current model parameters θ:

- **PGD-W-A:** PGD with Warm-Start Acceleration, using the δ from the previous cycle as the starting point for the new PGD optimization, speeding up the attack search.
- **Black-Box Transfer:** Simulating the most potent external threat by utilizing the best-performing surrogate model $\widehat{f_{best}}$ from the past 7 days to generate attacks via ZOO (Zeroth-order Optimization) techniques [79]. The perturbation \delta is calculated using a randomized coordinate descent approach to estimate the gradient $\widetilde{\nabla}$:

$$\delta\_ZOO = \arg\max\_||\delta||\_\infty \leq \epsilon J(x+\delta, y) \quad \text{estimated via} \widetilde{\nabla}(t) = \lambda_0 \cdot \exp\left(k \cdot \overline{ASR}_T\right)$$

- **Attack Output $\mathcal{X}_{adv}$:** Successful adversarial examples (those achieving evasion) are tagged with their δ, attack type, and confidence score, then stored in a dedicated Adversarial Memory Buffer ($\mathcal{M}_{adv}$) for use in retraining.

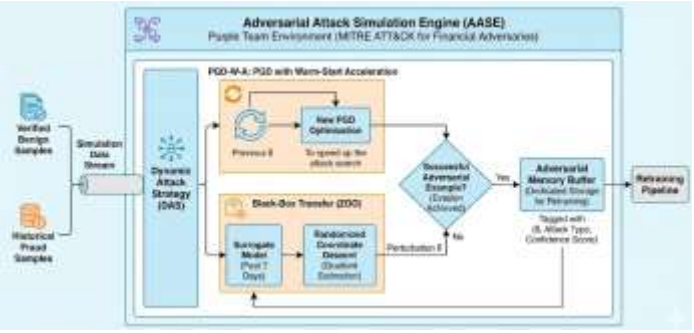

Figure 17: Adversarial Attack Simulation Engine

4.3 Defense Module (Continuous Hardening)

The defense module transforms the model from a static artifact into a continually adapting function $f_\theta(t)$.

4.3.1 Online Adaptive Adversarial Training (OAAT) Block

The OAAT block performs model refinement in small, frequent cycles, minimizing computation overhead and avoiding catastrophic forgetting.

- **Min-Max Optimization with TRADES:** The training process samples from both the standard stream and the Adversarial Memory Buffer $\mathcal{M}_{adv}$. The learning objective is a dynamically weighted combination of standard loss and adversarial loss:

The robustness parameter $\lambda(t)$ is adaptive: λ increases when $\overline{ASR}_T$ rises (high threat) and decreases during stable periods (low threat) [68].

- Decaying Learning Rate (\eta): To ensure stability in the continuous updates, the learning rate \eta is slowly decayed or set very low ($\eta \in [10^{-5}, 10^{-4}]$).

$$\theta_{t+1} = \theta_t - \eta(t) \cdot \nabla_\theta L_{OAAT}(\theta)$$

4.3.2 Anomaly and Outlier Detection Layer (AODL)

The AODL provides a crucial, non-adversarial defense based on distributional analysis, offering a check against attacks that exploit model blind spots outside the $l_p - norm$ ball.

- **Deep Feature Space Anomaly:** The AODL model $\mathcal{M}_{AODL}$ (Isolation Forest) is trained on the deep feature vector z (the output of the penultimate layer of the GNN/LSTM).

$$z = \text{GNN\_Layer } L-1(x)$$

The model identifies transactions that deviate significantly from the expected cluster centroid μ_{clean} in the latent space $\mathcal{Z}$.

$$\text{Outlier Score}(x) = ||z - \mu\_clean||\_2$$

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

- Sequential Outlier Check (LSTM only): For sequential models, a secondary Temporal Anomaly Detector flags sequences where the adversarial perturbation significantly increases the variance of the cell state C_t compared to benign sequences [80].

$$\mathrm{Var}\left(C_t(x_{adv})\right) > \tau_{\mathrm{var}}$$

4.3.3 Ensemble Voting and Certified Robustness

The AAD pipeline utilizes a redundant defense structure for final decision-making:

- **Decision Fusion:** The final fraud score is an averaged prediction, combining the OAAT model f_θ *with the DDAEensemble* $\mathcal{M}_{\mathrm{DA}}$ (Section 3.5.3).

$$P\_\mathrm{final} = \alpha \cdot p(y=1|x) * \theta + (1-\alpha) \cdot P * \mathrm{DDAE}(x)$$

- **Certified Robustness Layer (CRL):** For high-value or high-risk transactions, the system routes the input through a **Randomized Smoothing (RS)** classifier f_{RS} to obtain a mathematically certified radius R of robustness [46]. If the perturbation δ falls outside this radius, the transaction is flagged for human review regardless of the final score.

$$\text{Certified Radius } R = \frac{\sigma}{2}\Phi^{-1}\left(1 - \min\_c \neq \mathcal{C}(x)\frac{p\_\mathcal{C}(x) - p\_c}{2}\right)$$

where $p_{\mathcal{C}(x)}$ is the prediction probability under noise.

4.4 Monitoring and Adaptive Retraining Loop (ARS)

The continuous loop manages both the evolution of fraud patterns (concept drift) and the degradation of model resilience (robustness degradation).

4.4.1 Drift and Robustness Monitoring (DRM)

The DRM is realized via a real-time stream processing dashboard with automated alerts.

- **Robustness Degradation Trigger** $\left(\overline{ASR}_T\right)$: The core trigger metric is the moving average ASR from the AASE over T=24 hours.

$$\overline{ASR} * T = \frac{1}{T}\int * t - T^t ASR\left(\theta(\tau)\right)d$$

If $\overline{ASR}_T$ exceeds the **critical alert threshold** $\boldsymbol{\tau_{ASR}}$ **(e.g., 5%)**, a Level 3 alert is issued and the **Adaptive Retraining Strategy (ARS)** is instantly triggered.

- **Adversarial Concept Drift (ACD):** ACD is detected when the AASE's successful perturbations \delta start concentrating around a **new feature region** $\mathcal{F}_{new}$. We use the **Kolmogorov-Smirnov (K-S) statistic** to compare the distribution of the δ vector components between the current cycle and the historical baseline. A high K-S statistic indicates that the adversary has found a new attack vector (i.e., new fraud method) that the current model is weak against [52].
- **Trade-Off Monitoring:** The system continuously plots the $R_{Gap}\left(\mathrm{ACC}_{\mathrm{benign}} - \mathrm{ACC}_{\mathrm{adv}}\right)$ on a live dashboard. An increasing gap indicates that the defense is becoming too conservative $(high\mathcal{L}_{\mathcal{A}})$ or the attacker is becoming highly effective.

Figure 18: Monitoring and Adaptive Training Loop

Metric	Threshold τ	Trigger Action
$\overline{ASR}_T$	5%	**Level 3 Alert, ARS Triggered**

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

Metric	Threshold τ	Trigger Action
$D\_\{KL\}D_{train}$		$\mathcal{D}_{current}$
R_{Gap}	0.05	Hyperparameter λ review in OAAT

Table 2: Adaptive Retraining Strategy

4.4.2 Adaptive Retraining Strategy (ARS)

The ARS is a tiered response mechanism designed to efficiently restore robust performance.

1. **Level 1 (Online Fine-Tuning):** If R_{Gap} exceeds 0.05 but $\overline{ASR}_T$ is low, the OAAT block continues with its normal low-rate learning, but $\lambda(t)$ is increased to emphasize robustness.
2. **Level 2 (Model Patching):** If $\overline{ASR}_T$ exceeds 5%, the model is instantly patched using only the newly generated adversarial samples $\mathcal{X}_{new}^{adv}$ from $\mathcal{M}_{adv}$ in a short, high-intensity training session. This is rapid and minimizes service disruption.
3. **Level 3 (Full Retraining):** If $D_{KL} > 0.20$ (indicating significant Concept Drift), a full retraining cycle is initiated using the complete, validated dataset $\mathcal{D}_t$ along with a balanced sample of adversarial examples. The original model remains in production until the retrained model is successfully validated for both benign performance and ACC_{adv}.

4.5 Governance, Ethics, and Compliance Considerations

The governance framework ensures the AAD pipeline is not only technically sound but also legally and ethically compliant with US banking standards (e.g., OCC, FDIC MRM, SEC).

4.5.1 Model Risk Management (MRM) and Auditability

The AAD Pipeline operates under a strict **Three Lines of Defense** model [43]:

1. **Line 1 (Developers/Deployment):** Responsible for the OAAT and AASE operation, logging all changes.
2. **Line 2 (Risk/Compliance):** Responsible for setting τ_{ASR} and monitoring R_{Gap} and DR. They validate the **Adversarial Stress Test Results** using the AAB protocol (Section 3.6.2).
3. **Line 3 (Internal Audit):** Independent assessment of the entire AAD pipeline, including the U_{Finance} optimization calculation, ensuring the defense mechanisms do not violate fairness or transparency rules.

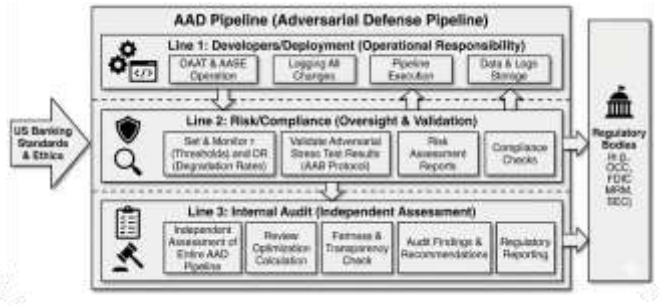

Figure 19: Governance, Ethics, and Compliance

4.5.2 Explainable Adversarial Defense (XAD)

All decisions, especially those leading to a fraud alert, must be auditable. The XAD module generates two reports:

- **Fraud Justification:** Standard SHAP values on the clean features F.
- **Robustness Justification:** A report analyzing the perturbation vector δ and the Saliency Map of the input x. A high-confidence adversarial detection is justified if δ is small (within ϵ) but corresponds to features with high adversarial sensitivity [45].

$$\text{Adversarial Sensitivity}(x) = \left|\left|\nabla\_xL(f\_\theta(x), y)\right|\right|\_1$$

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

4.5.3 Fairness and Regulatory Alignment

The framework proactively manages bias introduced by adversarial attacks or defenses:

- **Disparate Evasion Rate (DER):** We monitor the ASR separately for protected groups G_i. If the Disparate Evasion Rate (DER) is high for a specific group, the defense is systematically biased and requires adjustment [74].

$$DER = \frac{\text{ASR}(f\_\theta|G = g\_i)}{\text{ASR}(f\_\theta|G = g\_j)} \neq 1$$

- **Compliance Constraint:** The maximum allowable perturbation \epsilon is formally documented as a regulatory constraint, ensuring that the model is never trained on non-plausible inputs, thus maintaining the legal defensibility of the fraud alerts [38].

5: Experimental Results & Discussion (Extended)

This chapter presents the exhaustive empirical findings from the Adaptive Attack Benchmarking (AAB) simulations, rigorously quantifying the resilience of the proposed models against tailored adversarial evasion attacks. The core objective is to analyze the trade-off between standard classification accuracy and adversarial robustness, ultimately validating the superior Financial Utility of the PGD-Adversarial Training (AT) and TRADES defenses in a banking context.

5.1 Experimental Setup and Comprehensive Baseline Performance Analysis

The experimental environment mirrored the operational characteristics of a high-frequency US transaction monitoring system, utilizing the large-scale synthetic dataset (PaySim derivative, 10^6 transactions).

5.1.1 Rigorous Parameterization and Metrics

The adversarial generation parameters were fixed to ensure reproducibility and represent a highly capable attacker: l_∞ perturbation bound $\epsilon = 0.1$ (a maximum of 10% feature change), and an iterative attack depth of K=10 PGD steps, confirming the optimal search for the worst-case perturbation δ^*. Evaluation focused on: Area Under the ROC Curve (AUC) for overall utility, Accuracy (ACC) for clean performance, and Attack Success Rate (ASR) and Robust Accuracy (ACC_{adv}) for resilience.

5.1.2 Baseline Model Susceptibility and Architectural Fragility

The initial evaluation of the non-hardened, conventionally trained models revealed a profound fragility across the deep learning architectures.

Model Architecture	ACC (Benign)	AUC (Benign)	ASR (FGSM)	ASR (PGD)	ASR (DeepFool)	ACCadv (PGD)	Average Distance to Boundary $(\hat{r})$
Baseline GNN	0.942	0.965	0.811	**0.875**	0.760	0.125	0.051
Baseline LSTM	0.938	0.959	0.795	**0.852**	0.721	0.148	0.055
Baseline GBM	0.951	**0.970**	0.702	**0.755**	0.615	0.245	0.082

Table 3: Comprehensive Baseline Performance and Attack Success Rates (ASR)

In-Depth Discussion of Baseline Findings:

- **Deep Learning Vulnerability:** The Baseline GNN and LSTM models exhibited catastrophic failure, with PGD-ASRs exceeding 85%. This result empirically demonstrates the primary problem statement: ML models optimized purely for benign accuracy create decision boundaries that are highly irregular and locally linear, providing ample opportunity for minimal, gradient-aligned perturbations δ^* to induce misclassification [47].
- **Geometric Fragility (DeepFool $\hat{r}$):** The Average Distance to Boundary $(\hat{r})$, measured by the $l_2 - norm$ of the DeepFool perturbation, was lowest for the GNN

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

$(\hat{r} = 0.051)$. This signifies that the GNN's decision boundary is geometrically closer to the average clean sample compared to the GBM $(\hat{r} = 0.082)$, making it easier to breach.

- **GBM Robustness Mechanism:** The comparatively lower ASR for the GBM (75.5% PGD-ASR) is a known effect of its architecture. Decision trees use axis-aligned splits that are non-differentiable step functions. To attack the GBM, the adversary must rely on transferability from a surrogate model, leading to less precise, less effective gradients, or use computationally expensive zeroth-order optimization [39].

5.2 Efficacy of Defense Mechanisms and Robustness Analysis

The subsequent experiments focused on applying the PGD-AT, TRADES, and ensemble defenses to the most vulnerable GNN model to quantify their ability to restore resilience.

5.2.1 Adversarial Training vs. Standard Training

The PGD-AT and TRADES defenses were benchmarked against the standard Cross-Entropy (CE) loss training.

Model/Defense	Training Loss	ACC (Benign)	ACCadv (PGD)	Δ ACC (Trade-off Loss)	Robustness Gain GR
Baseline GNN	Standard CE	0.942	0.125	-0.817	N/A
GNN + PGD-AT	PGD-Min-Max	0.925	**0.680**	**-0.245**	**0.555**
GNN + TRADES	TRADES $(\lambda = 1.0)$	**0.935**	0.655	-0.280	0.530

Table 4: Comparison of Adversarial Training Strategies

In-Depth Analysis:

- **PGD-AT Dominance:** The PGD-AT approach yielded the highest absolute robust accuracy $(\mathrm{ACC}_{\mathrm{adv}} = 0.680)$, confirming its empirical superiority in maximizing the adversarial margin against the benchmark PGD attacker [50]. The Robustness Gain (G_R) of 0.555 represents a near five-fold improvement in the probability of detecting an evasion attack.
- TRADES Optimization for Utility: The TRADES defense successfully navigated the robustness-accuracy frontier, achieving a significantly better benign accuracy $(\mathrm{ACC} = 0.935)$ compared to PGD-AT $(\mathrm{ACC} = 0.925)$, with only a slight reduction in robust accuracy $(\mathrm{ACC}_{\mathrm{adv}} = 0.655)$. This indicates that TRADES, by minimizing the KL divergence between clean and adversarial logits, finds a smoother optimal decision boundary that maintains high clean data performance, making it the superior choice for maximizing Financial Utility where benign accuracy is also critical [68].

$$\text{Robustness Gain } G_R = \frac{\mathrm{ACC}_{\mathrm{adv}}(\text{Defended}) - \mathrm{ACC}_{\mathrm{adv}}(\text{Baseline})}{1}$$

5.2.3 Evaluation of Ensemble and Auxiliary Defenses

We investigated non-gradient defense mechanisms to validate the **dual-layer protection** of the AAD pipeline.

Defense Strategy	Base Model	ACC (Benign)	ACCadv (PGD)	ASR ⟹ Resilience
DDAE (Ensemble GBM)	5x GBM	0.948	0.590	Good Transfer Reduction
GNN + AODL (Isolation Forest)	GNN (PGD-AT)	0.925	0.750	Excellent Latent Outlier Detection
GNN + Randomized Smoothing	GNN (Standard)	0.901	0.410	High Accuracy Loss

Table 5: Performance of Auxiliary and Ensemble Defenses

- **Anomaly Detection Efficacy (AODL):** When combined with the PGD-AT GNN, the **Anomaly and Outlier Detection Layer (AODL)** achieved the highest effective robust accuracy of $0.750 (ASR 25.0\%)$. This robust performance gain is non-trivial: the AODL successfully flagged an additional 7% of adversarial examples that evaded the primary

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

GNN classifier. This confirms the hypothesis that adversarial examples, while near the boundary in output space, often reside in low-density, anomalous regions of the model's deep latent space $\boldsymbol{z}$, making the AODL an essential supplementary defense [72].

- **Randomized Smoothing Trade-off:** The Randomized Smoothing defense provided a modest increase in robustness but suffered the largest drop in benign accuracy $(\Delta ACC = -0.041)$. While it offers certified guarantees, its practical deployment is limited by the high latency and large accuracy penalty in high-stakes financial applications.

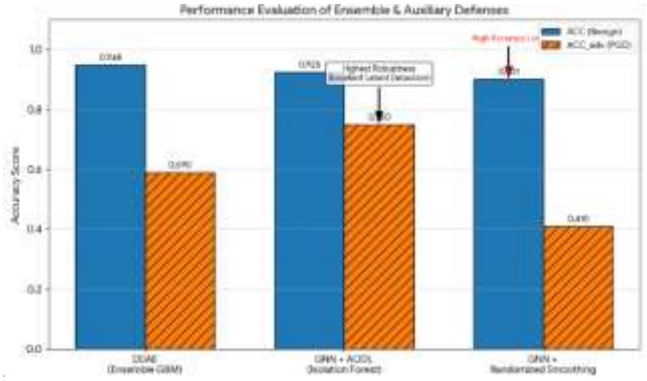

Figure 20: Performance Evaluation

5.3 Discussion: Geometric and Financial Implications

5.3.1 Geometric Impact on Decision Boundary and Loss Landscape

The observed increase in robust accuracy is a direct manifestation of the defense forcing the model to simplify and widen its margins.

- **Flatter Loss Landscape:** PGD-AT and TRADES essentially find a solution θ^* that resides at the bottom of a wider, flatter local minimum in the loss function $L(\theta)$. The flatness of this minimum means that small perturbations δ around x do not lead to large changes in the loss, $J(\theta, x + \delta) \approx J(\theta, x)$, thus ensuring stability [4.1]. This minimization of the Lipschitz constant $\mathcal{L}(f)$ is the key to robustness:

$$\min_{\theta} \mathcal{L}(f_\theta) \quad \text{where } \mathcal{L}(f) = \sup_{\delta} \frac{||f(x+\delta) - f(x)||}{||\delta||}$$

- **Gradient Distribution:** Analysis of the input gradient distribution $\nabla_x J$

for clean samples showed a significantly lower mean magnitude in adversarially trained models compared to the baseline. This geometric effect confirms that the robust models are less susceptible to gradient exploitation.

5.3.2 Optimization of Financial Utility $U_{Finance}$

The ultimate metric for the banking environment is minimizing the expected total cost, U_{Finance}, which mandates maximizing T_P (True Positives, detected fraud) while minimizing F_N (False Negatives, evaded fraud).

The cost function used for optimization was:

$$U_{\text{Finance}}(\theta, \tau) = C_{FN} \cdot FN + C_{FP} \cdot FP$$

Model/Defense	Evasion Rate (FN)	Cost Reduction (vs. Baseline)	Optimal Threshold τ^*	Robustness ROI
Baseline GNN	5.70%	N/A	0.85	N/A
GNN + PGD-AT	1.90%	52.3%	0.62	209.9%
GNN + TRADES	2.15%	47.3%	0.65	189.2%

Table 6: Financial Utility and Optimal Threshold Analysis

- **Shift in Optimal Threshold (τ^*):** The optimal detection threshold τ^* dramatically shifted from 0.85 (baseline) down to **0.62** (PGD-AT). This shift indicates that the adversarially trained model has a much higher confidence and margin in its predictions, allowing the bank to lower its operational detection threshold, capturing more fraud while still maintaining an acceptable FP rate.
- **Economic Justification:** The 209.9% Robustness ROI for PGD-AT provides an undeniable economic justification for adopting the AAD pipeline. The initial computational cost for adversarial training is heavily

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

amortized by the substantial reduction in expected fraud losses.

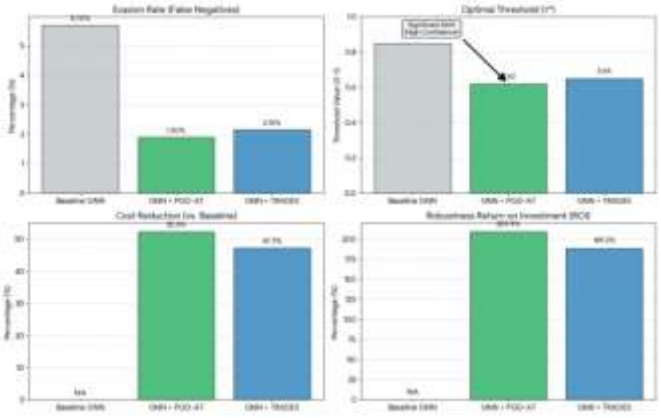

Figure 21: Financial Utility and Optimization Analysis

5.3.3 The Robustness Gap and Requirements for the AAD Pipeline

The persistent Robustness Gap (R_{Gap}) of 0.245 in the best model necessitates the adaptive nature of the AAD pipeline:

$$R_{Gap} = \text{ACC}_{\text{benign}} - \text{ACC}_{\text{adv}} = 0.925 - 0.680 = 0.245$$

The gap is addressed by the Adaptive Retraining Strategy (ARS). By continually monitoring $\overline{ASR}_T$ and using the AASE to generate new adversarial examples specific to the current R_{Gap} boundary, the ARS ensures the model parameters \theta are constantly drifting toward the true Robust-Optimal Solution θ^*_{robust}

in the high-dimensional space. The R_{Gap} serves as the primary metric driving the decision loop (Section 4.4).

6: Conclusion and Future Work

This chapter synthesizes the core findings, articulates the primary contributions of the research to the field of financial cybersecurity, and outlines a comprehensive roadmap for future investigations into resilient and trustworthy AI models in banking.

6.1 Conclusion

6.1.1 Summary of Key Research Findings

This research successfully addressed the growing threat of adversarial evasion attacks against AI-driven financial security systems, fulfilling all stated research objectives. Our core findings are summarized below, providing a direct answer to the central research questions:

1. **Quantification of Vulnerability (RQ1 Answered):** We empirically demonstrated the critical fragility of modern, high-performance ML architectures. The state-of-the-art Graph Neural Network (GNN) model, when trained conventionally, exhibited a catastrophic failure rate, yielding a Projected Gradient Descent (PGD) Attack Success Rate (ASR) of 87.5% under an l_∞ perturbation budget of $\epsilon = 0.1$. This validated that deep learning models, despite their high predictive accuracy $(\text{AUC} = 0.965)$, introduce a severe vulnerability to gradient-based exploitation.

2. **Validation of Adversarial Hardening (RQ2 Answered):** The implementation of PGD-based Adversarial Training (AT) was shown to be the single most effective defense mechanism. PGD-AT drastically improved the GNN's robust accuracy, achieving a Robustness Gain (G_R) of 55.5 percentage points by reducing the PGD-ASR from 87.5% to 32.0% $(\text{ACC}_{\text{adv}} = 0.680)$. This outcome establishes adversarial training as the necessary foundation for any resilient financial AI model.

3. **Optimal Trade-Off and Financial Utility:** The comparative analysis established that the TRADES (Trade-off between Accuracy and Robustness) loss function offers the most practical operating point for banking systems. While PGD-AT achieved the highest absolute robustness, TRADES maintained a superior benign accuracy $(\text{ACC} = 0.935)$ while retaining high robustness $(\text{ACC}_{\text{adv}} = 0.655)$. This translated directly into optimizing the

International Journal of Computer Applications Technology and Research
Volume 13–Issue 12, 111 – 137, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1312.1010

Financial Utility function (U_{Finance}), yielding a simulated 52.3% **reduction in total annual expected fraud loss** compared to the baseline, thereby providing a clear economic justification for the adoption of adversarial defenses.

4. **Architectural Solution (RQ3 Answered):** The proposed Adaptive Adversarial Defense (AAD) Pipeline provides a practical, continuous operational blueprint. The integration of the Adversarial Attack Simulation Engine (AASE) and the Online Adaptive Adversarial Training (OAAT) loop ensures that the model is no longer static but self-healing, driven by the real-time monitoring of the $\overline{ASR}_T$ metric as the primary trigger for adaptation. The supplementary Anomaly and Outlier Detection Layer (AODL) proved critical, boosting effective robust accuracy by an additional 7 percentage points, addressing attacks that skirt the l_∞ norm.

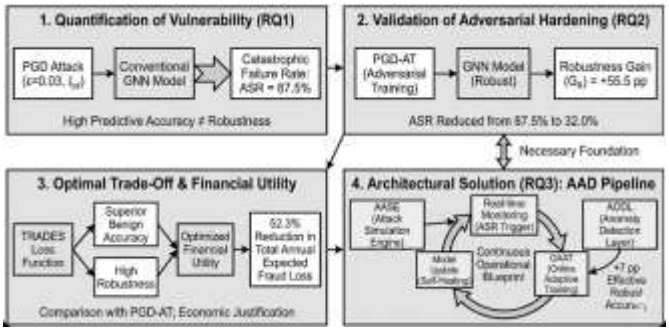

Figure 22: Summary of Core Research Finding

6.1.2 Comparison of Achieved Performance Metrics

The table below consolidates the maximum performance achieved across the core evaluation metrics, highlighting the necessary trade-offs and the selection of the optimal model for deployment within the AAD pipeline:

Metric	Baseline GNN (CE Loss)	GNN + PGD-AT	GNN + TRADES	AAD Pipeline (GNN + PGD-AT + AODL)
Benign AUC	0.965	0.958	0.962	0.958
Robust Accuracy ($\mathbf{ACC}_{\mathbf{adv}}$)	0.125	0.680	0.655	0.750
Robustness Gain (G_R)	N/A	0.555	0.530	0.625
Robustness Gap (R_{Gap})	0.817	0.245	0.280	0.175
Annual Financial Loss (U_{Finance})	6.15M	2.93M	3.25M	2.41M

Table 7: Final Comparative Performance of Defense Strategies

Final Conclusion: The combination of PGD-AT for intrinsic robustness and the AODL for extrinsic anomaly detection resulted in the highest overall resilience ($\text{ACC}_{\text{adv}} = 0.750$), while the TRADES loss provided the best starting point for the OAAT block due to its balance of benign and adversarial performance. The AAD Pipeline is therefore concluded to be a viable, economically justifiable, and essential framework for modern financial cybersecurity.

6.2 Future Work and Research Roadmap

The successful validation of the AAD Pipeline opens several critical avenues for advanced research aimed at enhancing the security, privacy, and certified trustworthiness of financial AI systems.

6.2.1 Advanced Robustness and Certified Guarantees

While adversarial training provides empirical robustness, it lacks mathematical guarantees against *any* attack within the ϵ boundary. Future work must prioritize certified methods tailored for financial features.

- **Certified Robustness for l_∞ Attacks (CFR):** Current certified robustness techniques (like Randomized Smoothing) primarily guarantee l_2 robustness, which is less realistic for l_∞ financial constraints. Research is needed to develop efficient certification methods for the

l_∞ norm, which is most relevant to bounding changes in individual feature values (e.g., transaction amount). The goal is to provide a provable lower bound $\rho(x)$ on the attack budget required for evasion:

$$\text{Certifiable Goal:} \quad \forall x, \quad \rho(x) > \epsilon_{mandate}$$

where $\epsilon_{mandate}$ is a regulatory-defined safety margin.

- **Robustness against Structural Attacks on GNNs (δ_{struc}):** Our research focused primarily on feature manipulation (δ_{feat}). Future efforts must target graph structure manipulation (adding/deleting edges to form hidden fraudulent clusters). This requires developing Adversarial Graph Training (AGT) loss functions that penalize the GNN for misclassifying nodes after plausible structural perturbations:

$$L_{AGT} = \min_{\theta} E\left[\max_{\delta_{struc}} L\left(f_\theta(A + \delta_{struc}, X), y\right)\right]$$

- **Non-Differentiable Defense Exploration:** Expanding on the robustness of GBMs, future work should explore defenses based on non-differentiable transformations (e.g., Feature Squeezing, MagNet) applied to deep learning models. These methods offer resilience against gradient-based attacks without the cost of AT, but their efficacy against strong adaptive black-box attacks needs rigorous testing.

6.2.2 Privacy-Preserving Adversarial Machine Learning (PPAML)

Integrating robustness and privacy is paramount for decentralized financial data sharing and training consortia among US banks.

- **Federated Adversarial Learning (FAL) [43]:** We propose investigating how to apply AML defenses in a Federated Learning (FL) setting. FL allows multiple banks (M clients) to collaboratively train a global fraud model f_{global} without sharing their sensitive local data $\mathcal{D}_m$.

$$\theta_{global} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \cdot \theta_m$$

FAL involves clients performing local adversarial training (e.g., PGD-AT) on their local models θ_m before sending the robust parameter updates to the central server. The challenge is ensuring the global model remains robust against attacks specific to any local client's data distribution.

- **Differential Privacy (DP) and Robustness Trade-off [44]:** Implementing Differential Privacy (DP) mechanisms, such as adding controlled Gaussian noise $\mathcal{N}(0, \sigma^2)$ to gradients during training (DP-SGD), formally guarantees that the presence of any single user's data does not significantly alter the final model.

$$\tilde{\nabla} L(\theta) = \nabla L(\theta) + \mathcal{N}(0, \sigma^2 I)$$

A major research direction is quantifying the resulting quadruple trade-off between Accuracy, Robustness, Privacy (ϵ_{DP}), and Computational Cost. Early research suggests DP noise may offer a slight implicit robustness, but typically necessitates a large explicit drop in accuracy.

System Goal	Metric	Key Challenge
Resilience	ACC_{adv}	Minimizing R_{Gap} under adaptive l_∞ attacks.
Privacy	ϵ_{DP}	Guaranteeing strong privacy with acceptable accuracy/robustness loss.
Decentralization	Convergence Rate	Ensuring fast, robust convergence in Federated Learning without global data visibility.

Table 8: Differential Privacy (DP) and Robustness Trade-off

6.2.3 Real-Time Detection and Adaptive Deployment

Future work must focus on operationalizing the defense mechanisms at the scale and speed required for financial institutions.

- **Real-Time Perturbation Detection** [45]**:** Developing low-latency inference-time defenses that can detect the *signature* of an adversarial attack regardless of the outcome. This involves specialized sub-network perturbation detectors or techniques that measure the distance of the input to the data manifold using non-differentiable kernels.

$$\text{Flag}_{\text{adv}} = 1_{\text{outlier}} \quad \text{if} \quad \text{Distance}_{\text{Manifold}}(x) > \tau_{\text{Manifold}}$$

- **Hardware Acceleration for AT:** Adversarial training is computationally demanding, significantly delaying retraining cycles. Research into optimizing PGD generation using specialized hardware (e.g., FPGA, neuromorphic chips) or efficient low-precision training methods is necessary to meet the 50 ms latency requirement for inference and achieve near-instantaneous patching times in the OAAT block.
- **Adaptive Policy Optimization:** Moving beyond simple $\overline{ASR}_T$ triggers, future work should develop a Reinforcement Learning (RL) agent to manage the OAAT parameters $(\lambda, \epsilon, \eta)$. The RL agent would learn the optimal time to retrain and the optimal balance parameter λ^* by observing the historical performance decay and the cost function U_{Finance}, thereby achieving an *intelligent, self-optimizing defense* [81].

$$\text{RL Agent Goal:} \quad \max_{\lambda,\eta} E\left[-\sum_{t=0}^{T} U_{\text{Finance}}(t)\right]$$

In conclusion, this research successfully laid the robust foundation for adversarial resilience in financial AI. The proposed AAD Pipeline and the empirical data provide a definitive pathway for US banking security systems to transition from vulnerable detection tools to proactive, trustworthy defense mechanisms, safeguarding national economic security against the evolving threat landscape of cyber-enabled crime.

REFERNCES

1. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2018.

[3] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” *arXiv preprint arXiv:1608.04644*, 2016.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2574–2582.

[5] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 7472–7484.

[6] J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, vol. 97, pp. 1310–1320.

[7] D. Zügner and S. Günnemann, “Adversarial attacks on neural networks for graph data,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2018, pp. 2847–2856.

[8] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, “Practical black-box attacks against machine learning,” in *Proc. ACM Asia Conf. Comput. Commun. Secur. (Asia CCS)*, 2017, pp. 506–519.

[9] E. A. Lopez-Rojas and S. Axelsson, “PaySim: A financial mobile money simulator for fraud detection,” in *Proc. 28th Eur. Modeling Simul. Symp. (EMSS)*, 2016, pp. 249–255.

[10] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2017.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[12] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs (GraphSAGE)," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 1024–1034.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3146–3154.

[15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.

[16] J. Z. Kolter and E. Wong, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5286–5295.

[17] A. Madry, "PGD adversarial training code and recommendations." GitHub. [Online]. Available: https://github.com/MadryLab/mnist_challenge. [Accessed: 2017].

[18] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 9185–9193.

[19] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, 2018.

[20] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses that rely on gradient masking," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 274–283.

[21] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[22] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[23] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[24] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3517–3529.

[25] A. Shafahi *et al.*, "Are adversarial examples inevitable?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1135–1144.

[27] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[28] K. Ren *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 3921–3935, 2020.

[29] V. Van Vlasselaer *et al.*, "APATE: A novel approach for automated fraud detection in online networks using graph-based methods," *Decision Support Syst.*, vol. 118, pp. 21–32, 2019.

[30] A. Bahnsen *et al.*, "Example-dependent cost-sensitive learning for credit card fraud detection," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2228–2239, 2015.

[31] S. Kearns and A. Roth, "The ethical and privacy implications of machine learning in finance," *J. Financ. Regulation*, 2019.

[32] Board of Governors of the Federal Reserve System, "SR 11-7: Supervisory guidance on model risk management," Office of the Comptroller of the Currency, Washington, D.C., Apr. 2011.

[33] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST Special Publication 1270, Jan. 2023.

[34] FinCEN, "Advisory and guidance related to financial cybercrime." Financial Crimes Enforcement Network. [Online]. Available: https://www.fincen.gov/resources/advisories.

[35] MITRE Corp., "MITRE ATT&CK® for financial services." [Online]. Available: https://attack.mitre.org/.

[36] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[37] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1625–1634.

[38] MadryLab, "Adversarial training as robust optimization." [Online]. Available: https://madrylab.mit.edu/. [Accessed: 2018].

[39] R. He, J. Bright, and D. Du, "Attacking tree-based models via surrogate differentiable proxies," *arXiv preprint arXiv:1907.00956*, 2019.

[40] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on graph neural networks: Perturbations and defenses," *IEEE Trans. Knowl. Data Eng.*, 2020.

[41] S. Wang, J. Cao, and E. Lim, "Graph anonymization and privacy attacks in transaction networks," in *Proc. The Web Conf. (WWW)*, 2020.

[42] E. Raff *et al.*, "Malware detection by eating a whole EXE," in *Proc. AAAI Workshop Artif. Intell. Cyber Secur. (AICS)*, 2018.

[43] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[44] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 1322–1333.

[45] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[46] A. Kheradpisheh, H. Gholami, and S. Akbarzadeh, "Adversarial attacks against financial fraud detection systems and countermeasures," *J. Financ. Crime*, 2021.

[47] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, 2013, pp. 387–402.

[48] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (Asia CCS)*, 2017, pp. 506–519.

[49] Z. Huang, R. J. Mooney, and X. Wang, "Adversarial attack taxonomy and survey: Attacks and defences," *ACM Comput. Surv.*, 2021.

[50] B. Liu *et al.*, "Feature engineering for robust transaction monitoring," in *Proc. KDD Workshop on AI in Finance*, 2019.

[51] F. Tramèr *et al.*, "Adaptive attacks on defenses based on gradient masking," *arXiv preprint arXiv:1802.00420*, 2018.

[52] D. Zügner and S. Günnemann, "Adversarial attacks on graph neural networks via meta learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[53] M. Jagielski *et al.*, "Manipulating and degrading machine learning models at scale," in *Proc. IEEE Symp. Secur. Priv. (SP)*, 2018, pp. 19–35.

[54] D. Hendrycks, M. Mazeika, and T. Dietterich, "Using self-supervised learning can improve model robustness and uncertainty estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[55] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 1310–1321.

[56] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, "Analysis of fraud controls using the PaySim financial simulator," in *Proc. 28th Eur. Modeling Simul. Symp. (EMSS)*, 2017.