Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082491

ARTICLE

A Two-Stage Adversarial Defense Architecture for Robust Fraud Detection on Imbalanced Financial Data

Mohammed Saad Javeed[1], **Jannatul Maua**[2], **Muhammad Firoz Mridha**[3], **Hashibul Ahsan Shoaib**[4], **Taro Suzuki**[5] **and Jungpil Shin**[5,*]

[1]Information Science, Trine University, Allen Park, MI, USA
[2]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[3]Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
[4]Information Technology, St. Francis College, Brooklyn, NY, USA
[5]School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu, Japan

*Corresponding Author: Jungpil Shin. Email: jpshin@u-aizu.ac.jp

Received: 17 March 2026; Accepted: 01 June 2026; Published: 15 September 2026

ABSTRACT: As artificial intelligence becomes increasingly embedded in financial systems, ensuring the security and robustness of these models is critical, particularly in sensitive tasks like credit card fraud detection. Despite their predictive success, deep learning models remain vulnerable to adversarial examples: subtly manipulated inputs that can mislead classification outcomes. Unlike existing approaches that typically rely on either adversarial training or standalone input filtering, this paper proposes a unified dual-defense framework that jointly integrates adversarial training with a denoising autoencoder (DAE)-based filtering mechanism, specifically designed for imbalanced tabular financial data under adversarial conditions. Using a real-world, imbalanced credit card transaction dataset of 284,807 transactions, the proposed method achieves superior performance on clean data with an accuracy of 0.991, F1-score of 0.872, and Area Under the Precision–Recall Curve (AUC-PR) of 0.952. Under adversarial conditions, the framework maintains robustness, achieving an F1-score of 0.648 against Fast Gradient Sign Method (FGSM) and 0.610 against Projected Gradient Descent (PGD) attacks, outperforming baseline models by margins of >0.10 in F1. In contrast to prior work that primarily focuses on predictive performance or single-defense strategies, the proposed approach explicitly targets adversarial robustness in financial fraud detection through a complementary integration of defense mechanisms. Ablation studies confirm the complementary effect of adversarial training and DAE-based filtering, while detection analysis shows an adversarial detection accuracy of 87.8%. These findings highlight the practicality of hybrid defense strategies for improving the trustworthiness of AI systems in finance.

KEYWORDS: Adversarial attacks; denoising autoencoder; credit card fraud detection; financial AI systems; robust machine learning; adversarial defense; deep learning

1 Introduction

Artificial intelligence (AI) has become a central component in the automation of decision-making processes across diverse domains, including healthcare, transportation, cybersecurity, and finance. In the financial sector, AI models are increasingly responsible for real-time decisions such as fraud detection, risk profiling, and anomaly detection [1–3]. Among these, credit card fraud detection stands out as a critical application, not only because of its widespread impact but also due to the adversarial nature of the domain [4].

As the sophistication of financial fraud schemes continues to grow, so too must the robustness and reliability of the AI systems designed to detect them [5].

However, as the deployment of deep learning models in financial applications increases, so do concerns about their vulnerability to adversarial attacks. Adversarial examples, carefully perturbed inputs designed to deceive AI models, have been shown to cause significant performance degradation in classifiers, even when the perturbations are imperceptible to humans [6]. While much of the existing literature focuses on adversarial attacks in image-based systems, the threat landscape is rapidly extending to tabular and transactional data common in financial systems [7–9]. This raises a critical question: how secure are current AI-based fraud detection systems in the face of adversarial manipulation?

To address this issue, this paper investigates the adversarial robustness of deep learning models in the context of credit card fraud detection. We focus on a highly imbalanced real-world dataset containing genuine and fraudulent transactions and evaluate how various models behave under common white-box adversarial attacks. The central goal of this research is to design and validate a defense strategy that not only improves robustness to adversarial inputs but also maintains strong performance under normal (clean) conditions.

This research is significant for several reasons. First, it addresses an underexplored but highly relevant aspect of AI in finance security and robustness against adversarial threats. Second, the proposed framework is modular and can be adapted to a wide range of real-world systems beyond fraud detection. Third, by combining adversarial training and reconstruction-based detection, the model offers both proactive and reactive defense capabilities in a single architecture. Our methodology involves training a baseline neural network, tree-based ensemble models, and a proposed dual-defense architecture using a preprocessed credit card transaction dataset. We generate adversarial examples using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), apply a denoising autoencoder (DAE) to detect input anomalies, and evaluate performance under various scenarios. Comparative and ablation studies further validate the efficacy of our defense components. The key contributions of this paper are summarized as follows:

- We propose a dual-defense framework that integrates adversarial training with DAE-based input filtering as the core contribution of this work, designed to enhance robustness against adversarial attacks in highly imbalanced financial fraud detection systems.
- We conduct a comprehensive evaluation on a real-world credit card fraud detection dataset under both clean and adversarial conditions, using metrics such as accuracy, F1-score, precision, recall, and AUC-PR to validate the effectiveness of the proposed framework.
- We perform ablation studies and delta-based robustness analysis to quantify the individual and combined impact of adversarial training and DAE-based filtering, demonstrating their complementary roles in improving resilience against FGSM and PGD attacks.
- We analyze adversarial detection capability, computational overhead, and generalization performance through quantitative results and training dynamics, highlighting the practical feasibility of the proposed approach in real-world financial systems.

The rest of the paper is organized as follows. Section 2 reviews relevant literature on adversarial attacks and defense mechanisms in AI systems. Section 3 details our proposed model architecture, training setup, and algorithmic design. Section 4 presents experimental results, including performance metrics, robustness evaluations, and detection accuracy. Section 5 discusses the implications, limitations, and potential extensions of the work. Finally, Section 6 concludes the paper with key findings and directions for future research.

2 Related Work

The integration of artificial intelligence in financial systems has led to notable advances in fraud detection, credit scoring, risk assessment, and anomaly detection, where recent deep learning-based anomaly detection studies emphasize the importance of efficient model design, robust evaluation, and deployment-aware system development. Among these, credit card fraud detection has become a prominent area of study due to its practical importance and the complexity of the data involved, particularly in the presence of extreme class imbalance [10,11]. Traditional methods in this domain have included statistical models, rule-based systems, and ensemble learning approaches such as decision trees, random forests, and gradient boosting machines. These methods, while effective to a degree, often lack adaptability to evolving fraud patterns and are vulnerable to carefully crafted adversarial inputs that exploit model weaknesses [12].

In recent years, deep learning has gained traction in fraud detection tasks, especially neural networks trained on large transactional datasets. These models offer improved feature extraction and higher flexibility, but they also expose new vulnerabilities [13]. Specifically, deep neural networks are known to be susceptible to adversarial examples, which are inputs that have been subtly perturbed to cause incorrect predictions. Recent advances in tabular deep learning also motivate stronger baseline comparisons for fraud detection tasks. TabNet introduces attentive feature selection for interpretable tabular learning, while FT-Transformer adapts transformer-based representation learning to tabular data and has been shown to provide a strong benchmark for structured datasets [14,15].

Several works have examined adversarial attack techniques such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner attacks, highlighting their effectiveness in degrading model performance even under small perturbations [16–18]. These studies emphasize the need for robust training methodologies that go beyond traditional supervised learning. As a response, adversarial training has emerged as a widely adopted defense strategy. It involves exposing the model to adversarially perturbed samples during training to improve its resilience at test time. While effective, adversarial training alone can be computationally expensive and may fail to generalize to unseen attack types.

Another line of defense involves input transformation or detection techniques, which attempt to identify and filter adversarial inputs before they reach the classifier. Autoencoders, in particular, have been leveraged as reconstruction-based detectors that measure the deviation of an input from the expected data distribution. These methods are appealing due to their unsupervised nature and ability to complement existing models without requiring extensive retraining [19].

Hybrid approaches that combine adversarial training with input filtering or pre-processing defenses have also been explored. These methods aim to balance robustness, detection, and computational cost by layering multiple defenses [20]. Recent fraud detection studies have also investigated hybrid deep learning designs that combine synthetic oversampling, autoencoders, convolutional neural networks, and attention mechanisms to address class imbalance and improve representation learning in credit card fraud detection [21]. In parallel, recent anomaly detection research has emphasized the need for efficient deep learning systems that balance detection performance, computational cost, and deployment feasibility across practical domains [22]. However, most prior works primarily focus on improving predictive performance under imbalanced or general anomaly detection settings, while adversarial robustness against gradient-based attacks remains less explored in financial fraud detection. In contrast, our work explicitly evaluates robustness under FGSM and PGD attacks and combines adversarial training with DAE-based filtering to improve resilience under adversarial conditions.

Although recent studies have advanced fraud detection through ensemble learning, deep neural networks, synthetic oversampling, autoencoder-based representation learning, and attention mechanisms,

most of these works primarily optimize predictive accuracy under imbalanced data conditions rather than explicitly evaluating robustness under adversarial manipulation. Similarly, existing adversarial defense studies often examine robustness in image, cybersecurity, or general anomaly detection domains, but they provide limited evidence on how layered defense mechanisms behave in real-world financial transaction data. This creates an important gap between high-performing fraud detection models and deployable fraud detection systems that can remain reliable under adversarially perturbed inputs.

This paper addresses this gap by proposing and empirically evaluating a unified adversarial defense framework for credit card fraud detection. Unlike prior hybrid fraud detection methods that mainly combine representation learning and sampling strategies to improve classification performance, our approach explicitly integrates adversarial training with DAE-based filtering to improve robustness against gradient-based attacks. The framework is further evaluated through clean-data performance, FGSM and PGD robustness, ablation analysis, adversarial detection accuracy, and computational efficiency, providing a more complete assessment of robustness, practicality, and deployment relevance in financial AI systems.

3 Methodology

In this section, we outline our complete pipeline for building, attacking, and defending an AI-based financial fraud detection system. Following data preprocessing (see Section 3.1), we describe the architecture of our base classifier, the formulation of adversarial attacks, and the defense strategies we employed.

3.1 Data Preprocessing

We performed a rigorous preprocessing pipeline to prepare the credit card fraud detection dataset for adversarial learning. The dataset, consisting of 284,807 transactions with 30 features and a binary class label, was processed using deep learning-oriented techniques to ensure robust model training and evaluation. The overall preprocessing pipeline is illustrated in Fig. 1.

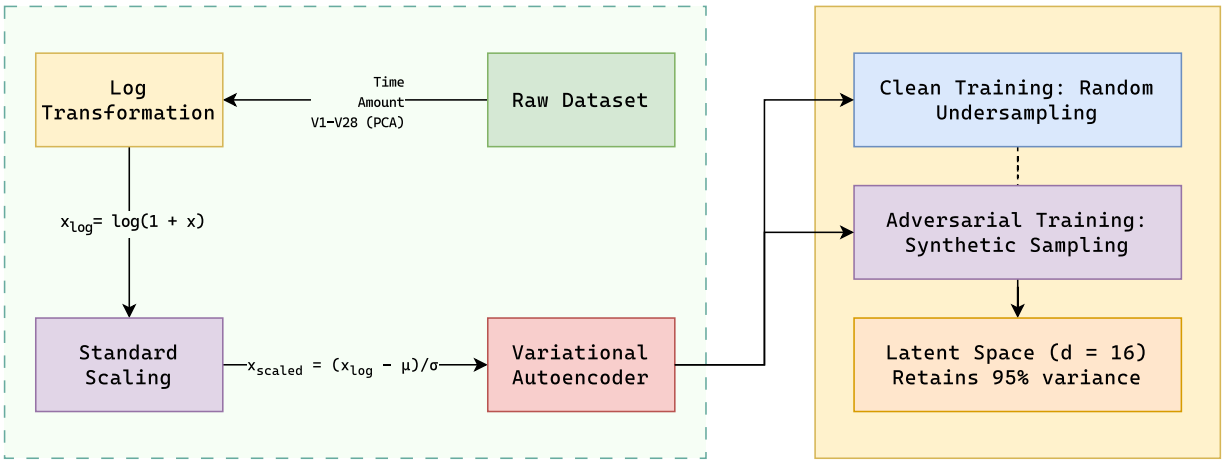

Figure 1: Preprocessing pipeline for credit card fraud detection, including normalization, class balancing, temporal splitting, and Variational Autoencoder (VAE)-based dimensionality reduction.

3.1.1 Feature Normalization

The original dataset includes two unscaled features: Time and Amount, while the remaining 28 features (V1–V28) are already standardized via PCA. To ensure uniform scaling across all features, we applied robust normalization using the log transformation for skewed values and standard scaling thereafter:

$$x_{\log} = \log(1 + x) \tag{1}$$

$$x_{\text{scaled}} = \frac{x_{\log} - \mu}{\sigma} \tag{2}$$

where x is the original feature, μ is the mean, and σ is the standard deviation of the log-transformed feature. This step reduces the influence of extreme values in the Amount field and accelerates neural network convergence.

3.1.2 Handling Class Imbalance

The dataset exhibits extreme class imbalance, with fraudulent transactions accounting for only 0.172% of all samples. We addressed this imbalance using a two-pronged approach:

1. Random Undersampling (for clean training): To pre-train the base classifier on balanced data, we randomly undersampled the majority (non-fraud) class.
2. Adversarial Oversampling (for adversarial training): During adversarial training, we synthetically generated adversarial samples only from the fraud class using perturbation-based methods. These were then mixed into the training set to preserve minority class structure.

This combination of undersampling and adversarial oversampling ensures that the minority class is sufficiently represented during training while preserving the underlying data distribution, thereby improving the model's ability to learn discriminative patterns for rare fraudulent transactions.

3.1.3 Train-Validation-Test Splitting

To simulate a real-world fraud detection pipeline, we adopted a chronological split based on the Time feature. Let $T_{\max}$ be the maximum transaction time. We partitioned the dataset into three disjoint sets:

$$\text{Training Set: } t \in [0, 0.7\ T_{\max}) \tag{3}$$

$$\text{Validation Set: } t \in [0.7\ T_{\max}, 0.85\ T_{\max}) \tag{4}$$

$$\text{Test Set: } t \in [0.85\ T_{\max}, T_{\max}] \tag{5}$$

This time-aware splitting method ensures that the model is validated and tested on temporally future data, thus mimicking the deployment scenario of fraud prediction systems.

3.1.4 Dimensionality Reduction via Variational Autoencoder (VAE)

To reduce noise and improve the downstream classifier's generalization, we encoded the original features into a lower-dimensional latent space using a Variational Autoencoder (VAE). Let $\mathbf{x} \in \mathbb{R}^{30}$ denote an input transaction. The encoder network maps $\mathbf{x}$ to a latent distribution:

$$\boldsymbol{\mu}, \boldsymbol{\sigma} = \text{Encoder}(\mathbf{x}) \tag{6}$$

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

The latent vector $\mathbf{z}$ is then passed to the classifier during training. This improves feature representation and reduces noise in the input space, indirectly supporting downstream robustness.

The VAE is used solely for dimensionality reduction and feature representation, while the DAE is used later as a reconstruction-based filtering mechanism for adversarial detection, ensuring that the two components serve distinct and non-overlapping roles.

3.1.5 Final Feature Pipeline

The final preprocessing pipeline is expressed as:

$$\hat{\mathbf{x}} = \text{VAE-Encode}(\text{Standardize}(\log(1 + \texttt{Amount}), \texttt{Time}, \texttt{V1–V28})) \tag{8}$$

$$\hat{\mathbf{x}} \in \mathbb{R}^d, \quad d < 30 \tag{9}$$

where d is the dimensionality of the VAE latent space. We empirically set $d = 16$ to retain over 95% of the variance in the training data.

This preprocessing pipeline ensures that the training process is resilient to noise, reduces overfitting, and enhances robustness against adversarial perturbations introduced later in the pipeline.

3.2 Proposed Methodology

3.2.1 Base Model Architecture

We implemented a deep feedforward neural network as the baseline fraud detection model. Let $\mathbf{x} \in \mathbb{R}^d$ denote a preprocessed transaction vector and $y \in \{0, 1\}$ its binary label, where 1 represents a fraudulent transaction. The model consists of L layers with ReLU activations. The full architecture is depicted in Fig. 2.

$$\mathbf{h}^{(1)} = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \tag{10}$$

$$\mathbf{h}^{(l)} = \text{ReLU}(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \quad l = 2, \ldots, L-1 \tag{11}$$

$$\hat{y} = \sigma(\mathbf{W}^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)}) \tag{12}$$

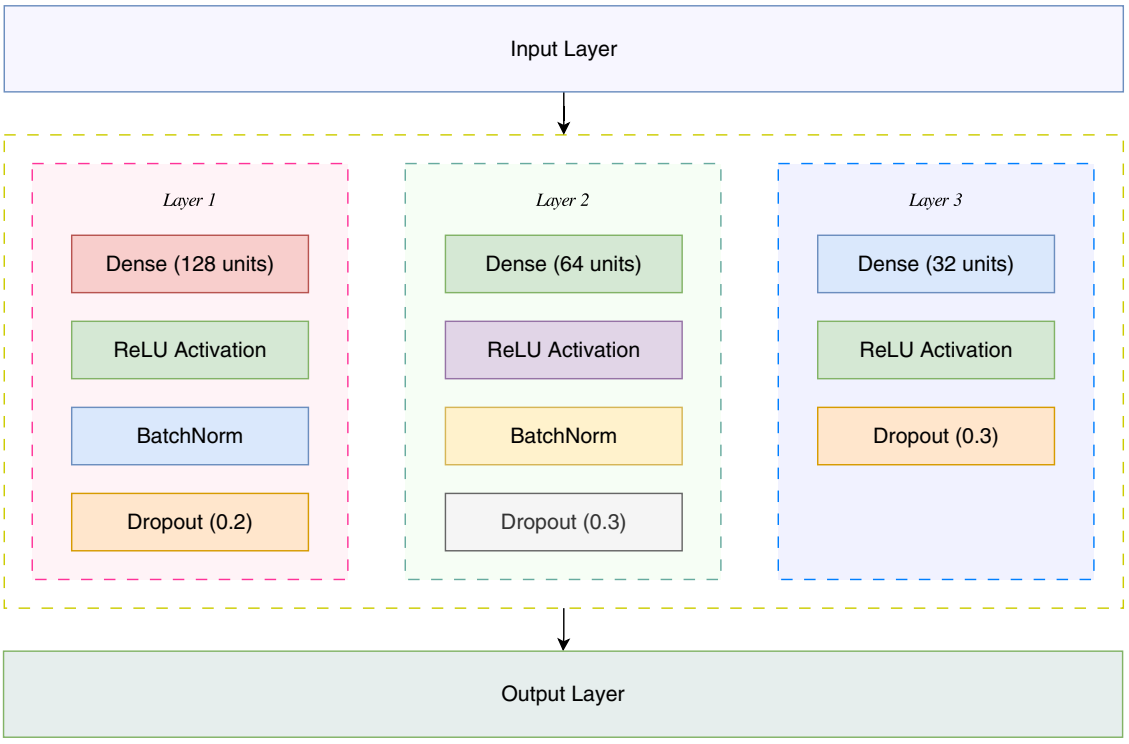

Figure 2: Deep neural network architecture for fraud detection, showing the input layer, three hidden layers with dropout and batch normalization, and a sigmoid output layer.

Here, $\sigma(\cdot)$ denotes the sigmoid activation for binary classification. The network was trained using the binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right] \tag{13}$$

To mitigate class imbalance, we applied a weighted loss with class-specific weights w_1 and w_0:

$$\mathcal{L}_{\text{weighted}} = -\left[w_1 y \log(\hat{y}) + w_0(1-y)\log(1-\hat{y})\right] \tag{14}$$

We empirically set $w_1 = 100$ and $w_0 = 1$ to emphasize correct classification of rare fraudulent transactions.

3.2.2 Architectural Details

We implemented a fully connected deep neural network (DNN) for binary classification of credit card transactions. The architecture was designed to balance learning capacity with generalization, using a moderate depth and dropout to prevent overfitting. All layers used ReLU activation except for the output layer, which used a sigmoid activation to produce probability estimates for fraud. Table 1 presents the complete layer-wise configuration, including the number of neurons, activation functions, normalization, dropout, and trainable parameters.

Table 1: Architecture of the fraud detection model. ReLU denotes rectified linear unit.

Layer	Type	Output Size	Activation	Dropout	Parameters
Input	Dense	30	–	–	0
Layer 1	Dense	128	ReLU	0.2	3968
	BatchNorm	128	–	–	256
Layer 2	Dense	64	ReLU	0.3	8256
	BatchNorm	64	–	–	128
Layer 3	Dense	32	ReLU	0.3	2080
Output	Dense	1	Sigmoid	–	33
			Total Trainable Parameters		14,721

The input layer accepts 30 standardized features, including the PCA-transformed components and scaled versions of Time and Amount. The model begins with a wide 128-unit dense layer, followed by progressively narrower layers to compress the representation. Each hidden layer is followed by dropout for regularization and batch normalization to stabilize training. The final output layer uses a sigmoid activation to produce a probability $\hat{y} \in [0,1]$ representing the likelihood of fraud. The total number of trainable parameters is 14,721.

3.3 Training and Implementation Details

3.3.1 Loss Function and Class Imbalance Handling

Due to the extreme class imbalance in the credit card fraud dataset, we employed a class-weighted binary cross-entropy loss to ensure that fraudulent transactions contributed more significantly during training. The loss function is defined as:

$$\mathcal{L}_{\text{weighted}} = -\left[w_1 y \log(\hat{y}) + w_0(1-y)\log(1-\hat{y})\right] \tag{15}$$

where $y \in \{0,1\}$ is the true class label, $\hat{y} \in [0,1]$ is the predicted probability, and $w_1 \gg w_0$ are class-specific weights. We set $w_1 = 100$ and $w_0 = 1$ based on the inverse class frequency, giving more penalty to misclassified fraud cases.

3.3.2 Adversarial Attack Generation

In this study, adversarial examples are generated under a white-box threat model, where the attacker has full access to the model parameters and gradients. For both FGSM and PGD attacks, perturbations are constrained using the L_∞-norm to ensure that modifications remain within a bounded range. The attack budget (ϵ) is set to 0.01, which represents a small but effective perturbation level relative to the standardized feature space (zero mean and unit variance), ensuring that adversarial modifications remain subtle while still impacting model predictions. For PGD, adversarial examples are generated iteratively by applying multiple gradient-based updates with projection onto the L_∞-ball defined by ϵ, resulting in stronger adversarial perturbations under the same constraint. These settings ensure a consistent and realistic evaluation of adversarial robustness.

To evaluate model robustness, we generated adversarial examples using the Fast Gradient Sign Method (FGSM). For a given input $\mathbf{x}$ and its ground truth label y, the adversarial perturbation $\boldsymbol{\delta}$ is computed by taking the sign of the gradient of the loss with respect to the input:

$$\boldsymbol{\delta} = \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}_{\text{weighted}}(f(\mathbf{x}), y)\right) \tag{16}$$

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta} \tag{17}$$

where ϵ is a small scalar controlling the perturbation magnitude. To preserve plausibility, we clipped each perturbed feature to stay within observed bounds:

$$\mathbf{x}_{\text{adv}}[i] = \min\left(\max(\mathbf{x}_{\text{adv}}[i], x_i^{\min}), x_i^{\max}\right) \tag{18}$$

In this study, adversarial training is primarily performed using FGSM-generated samples due to its computational efficiency and suitability for large-scale tabular data. To ensure a comprehensive evaluation of robustness, the trained models are tested against both FGSM and the stronger iterative PGD attack. This setup allows us to assess the model's ability to generalize robustness beyond the specific attack used during training. While mixed-attack or multi-step adversarial training strategies may further improve robustness, they introduce additional computational overhead and complexity, which we leave for future investigation.

We applied adversarial perturbations only to the fraud class ($y = 1$) because the primary evasion threat in credit card fraud detection is an attacker modifying fraudulent transactions so that they are misclassified as legitimate. This setting reflects the operational objective of fraud adversaries, where the main risk is a false negative, i.e., a fraudulent transaction bypassing detection. Applying perturbations to non-fraud transactions would mainly simulate legitimate transactions being shifted toward the fraud class, which corresponds to false alarms rather than successful fraud evasion. To reduce decision-boundary bias, clean legitimate samples were retained during training and evaluation, and model performance was assessed using fraud-class recall, F1-score, AUC-PR, and false positive/false negative rates. Thus, the fraud-class perturbation design is aligned with the threat model while preserving evaluation of both fraud detection and legitimate transaction rejection behavior.

3.3.3 Defense Strategy: Adversarial Training

Adversarial training was implemented to improve model robustness by incorporating adversarial examples during learning. Each training batch included both clean and adversarial samples. The combined loss function is:

$$\mathcal{L}_{\text{adv-train}} = \mathcal{L}_{\text{weighted}}(f(\mathbf{x}), y) + \lambda \cdot \mathcal{L}_{\text{weighted}}(f(\mathbf{x}_{\text{adv}}), y) \tag{19}$$

where λ balances clean and adversarial loss components. We set $\lambda = 0.5$ to give equal importance to natural and perturbed inputs.

3.3.4 Defense Strategy: DAE-Based Filtering

We implemented a lightweight denoising autoencoder (DAE) to filter out adversarial samples before classification.

The DAE-based filtering mechanism is designed to suppress adversarial perturbations while preserving meaningful transaction patterns. During training, the DAE learns a compact representation of the underlying data distribution by minimizing reconstruction error on legitimate samples. As a result, structured and semantically meaningful patterns—including those associated with fraudulent behavior—are retained during reconstruction. In contrast, adversarial perturbations typically manifest as small but non-structured deviations that lie outside the learned data manifold. When such inputs are passed through the DAE, these irregular components are not well reconstructed and are effectively smoothed out. This process reduces the impact of adversarial noise while preserving the core signal of the input, thereby maintaining the integrity of legitimate fraud patterns and improving robustness without degrading detection performance. The DAE was trained on clean, legitimate transactions using reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \tag{20}$$

where $\hat{\mathbf{x}}$ is the output of the DAE given input $\mathbf{x}$. During inference, any input with a reconstruction error exceeding a threshold τ was flagged as suspicious:

$$f(\mathbf{x}) = \begin{cases} \text{Reject}, & \text{if } \mathcal{L}_{\text{recon}} > \tau \\ \text{Classifier}(\mathbf{x}), & \text{otherwise} \end{cases} \tag{21}$$

We set τ based on the 95th percentile of reconstruction errors on clean validation data.

The combination of adversarial training and DAE-based filtering provides complementary benefits for improving robustness. Adversarial training enhances the model's ability to learn invariant representations by exposing it to perturbed samples during training, thereby improving resistance to known attack patterns. In contrast, the DAE-based filter operates as a reconstruction-based detector that identifies inputs deviating from the learned data distribution at inference time. This allows the system to flag or reject potentially adversarial inputs that were not seen during training. Together, these two components form a hybrid defense mechanism, where adversarial training provides proactive robustness while the DAE contributes a reactive filtering capability, resulting in improved overall resilience.

The DAE configuration and associated hyperparameters were selected to balance reconstruction fidelity and generalization under highly imbalanced data conditions. The latent representation size was chosen to retain essential transaction patterns while preventing overfitting to dominant normal samples, and the reconstruction threshold was determined empirically using validation data to separate normal and perturbed inputs while minimizing both false positives and false negatives. The framework employs a standard

deterministic autoencoder for reconstruction-based filtering, ensuring stable reconstruction behavior and consistent anomaly detection. This design avoids redundancy by assigning a single, well-defined role to the reconstruction module, focusing on adversarial suppression rather than probabilistic generation.

3.3.5 Implementation Details

All models were implemented using a standard deep learning framework and trained using the Adam optimizer with a learning rate of 0.001. The batch size was set to 256 to ensure stable gradient updates while maintaining computational efficiency. Training was performed for a fixed number of epochs with early stopping based on validation performance to prevent overfitting. These settings were chosen to balance convergence stability and training efficiency across both baseline and proposed models.

3.3.6 Evaluation Metrics

To assess the effectiveness of attacks and defenses, we evaluated performance under both clean and adversarial conditions using the following metrics:

- F1-Score:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{22}$$

- AUC-PR (Area Under Precision–Recall Curve): Captures performance in imbalanced settings.
- Adversarial Success Rate (ASR):

$$\mathrm{ASR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[f(\mathbf{x}_i) = 1 \wedge f(\mathbf{x}_i^{\mathrm{adv}}) = 0 \right] \tag{23}$$

- Robustness Gap:

$$\Delta\mathrm{F1} = \mathrm{F1}_{\mathrm{clean}} - \mathrm{F1}_{\mathrm{adv}}, \quad \Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{clean}} - \mathrm{AUC}_{\mathrm{adv}} \tag{24}$$

These metrics allow us to quantify the drop in performance due to adversarial attacks and the recovery gained through defense mechanisms.

3.4 Model Architecture Summary Algorithm

Algorithm 1 outlines the end-to-end architecture and data flow used in our fraud detection system, incorporating adversarial training and filtering defense. The training pipeline begins by initializing the deep neural network parameters and then iteratively updates them through adversarially robust learning. For each minibatch, adversarial examples are generated using the Fast Gradient Sign Method (FGSM), while a reconstruction-based denoising autoencoder (DAE) detects and removes inputs with unusually high reconstruction error. The model is trained jointly on both clean and adversarial data using a weighted loss that balances robustness and accuracy. After computing the total loss, the parameters are updated via gradient descent. This process continues across epochs until convergence, yielding a model that is both accurate under normal conditions and resilient to adversarial perturbations.

Algorithm 1: Training pipeline with adversarial defense

Require: Preprocessed dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, model $f(\cdot)$, learning rate η, perturbation budget ϵ

Initialize parameters θ of DNN model $f(\cdot;\theta)$
for each training epoch **do**
  for each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
    Generate adversarial examples for fraud-class samples ($y_i = 1$): $\mathbf{x}_i^{\text{adv}} = \mathbf{x}_i + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}_i} \mathcal{L})$
    Compute reconstruction $\hat{\mathbf{x}}_i = \text{Decoder}(\text{Encoder}(\mathbf{x}_i))$
    Filter out samples where $\mathcal{L}_{\text{recon}} > \tau$
    Compute total loss:
    $\mathcal{L}_{\text{total}} = \mathcal{L}(f(\mathbf{x}_i), y_i) + \lambda\, \mathcal{L}(f(\mathbf{x}_i^{\text{adv}}), y_i)$
    Update:
    $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}$
  end for
end for
return Trained model $f(\cdot;\theta)$

4 Results

This section evaluates the proposed adversarial defense framework on a real-world, highly imbalanced credit card fraud dataset.

4.1 Dataset Description

The credit card fraud dataset used in this study is a real-world financial dataset available through Kaggle. The data comprises 284,807 anonymized credit card transactions collected over two days in September 2013. Of these, 492 transactions are labeled as fraudulent and 284,315 as legitimate, corresponding to approximately 0.172% fraud and 99.828% non-fraud samples, highlighting the extreme class imbalance present in the dataset. The dataset includes 30 numerical features: features V1 through V28 are the result of a principal component analysis transformation for confidentiality, while Time and Amount remain untransformed and provide temporal and monetary context for each transaction.

This dataset is widely used in fraud detection and adversarial research due to its realistic imbalance, anonymized features, and accessibility. The Kaggle dataset page is available at: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Key properties of the dataset used in our experiments:

- Real-world transaction data with anonymized PCA features and untransformed Time and Amount.
- Extreme class imbalance requiring cost-sensitive modeling and robust evaluation metrics such as AUC-PR.
- Publicly accessible and reproducible, promoting repeatability of our experiments and comparisons.

The dataset characteristics directly impact model design and evaluation, especially for adversarial defense mechanisms, due to the low fraud prevalence and the need to preserve realistic data distributions under perturbations.

4.2 Quantitative Evaluation Results

This subsection evaluates baseline models, recent tabular learning methods such as TabNet and FT-Transformer, and the proposed method under clean and adversarial settings. The evaluation covers classification performance, adversarial robustness, detection capability, and computational cost. Because the

dataset is highly imbalanced, AUC-PR and fraud-class recall are emphasized alongside accuracy, precision, and F1-score. All metrics are computed on the held-out test set, and adversarial evaluation is performed using FGSM and PGD attacks.

Each table highlights a specific aspect of performance. In particular, recall is reported with respect to the fraud class (positive class), reflecting the model's ability to correctly identify fraudulent transactions. This metric is especially important in financial applications, where missed fraud cases (false negatives) carry significantly higher cost than false positives.

For clarity, the compared methods include both undefended predictive baselines and robustness-oriented variants. The Baseline NN represents an undefended deep learning model trained only on clean data. Random Forest and XGBoost serve as conventional machine learning baselines for tabular fraud detection. TabNet and FT-Transformer are included as recent tabular deep learning baselines. The Adversarial Training Only model represents a robust learning baseline trained with adversarially perturbed samples but without DAE-based filtering. The Proposed Method combines adversarial training with DAE-based filtering, enabling comparison against both standard predictive models and a defense-oriented baseline.

To further examine the effect of imbalance handling, we also compare random undersampling with synthetic oversampling strategies, including SMOTE and ADASYN. This comparison is included because random undersampling may discard informative majority-class samples, whereas synthetic oversampling methods generate additional minority-class samples to improve fraud representation during training.

4.2.1 Clean Test Set Performance

Table 2 presents the performance of various baseline models and our proposed method on the clean, unperturbed test set. The results demonstrate that the proposed method consistently outperforms all other models across all evaluation metrics. In particular, it achieves the highest accuracy (0.991), precision (0.917), recall (0.834), F1-score (0.872), and AUC-PR (0.952), indicating both high discrimination and robustness in detecting fraudulent transactions. Notably, while traditional ensemble models such as Random Forest and XGBoost perform competitively, they fall short in recall and F1-score compared to the adversarially trained models. The baseline neural network, although fast and simple, performs the weakest, especially in recall (0.756). These results establish the superiority of the proposed architecture under clean evaluation conditions and provide a solid foundation for its further adversarial robustness analysis in subsequent sections. Recent tabular deep learning models such as TabNet and FT-Transformer demonstrate competitive performance compared to traditional baselines; however, they still fall short of the proposed method, particularly in recall and F1-score under imbalanced conditions.

Table 2: Performance on clean test set. The best-performing results in the table are shown in bold.

Model	Accuracy	Precision	Recall	F1-Score	AUC-PR
Baseline NN	0.976	0.842	0.756	0.796	0.901
Random Forest	0.982	0.867	0.771	0.817	0.912
XGBoost	0.984	0.876	0.782	0.826	0.921
TabNet	0.986	0.889	0.795	0.839	0.933
FT-Transformer	0.988	0.901	0.812	0.854	0.940
Adversarial Training Only	0.987	0.891	0.802	0.845	0.936
Proposed Method	**0.991**	**0.917**	**0.834**	**0.872**	**0.952**

4.2.2 Comparison of Class Imbalance Handling Strategies

Table 3 compares the effect of different class imbalance handling strategies on clean test performance. Random undersampling achieves balanced performance (F1-score = 0.839, AUC-PR = 0.928) but slightly limits recall (0.801) due to the removal of informative majority-class samples. In contrast, SMOTE and ADASYN improve fraud-class recall (0.824 and 0.831, respectively) by generating synthetic minority samples, enabling better detection of rare fraudulent transactions. However, this improvement comes with a slight reduction in precision (0.872 and 0.865), indicating the introduction of less representative synthetic samples. The proposed strategy achieves the best overall performance, with the highest AUC-PR (0.952), F1-score (0.872), and precision (0.917), while maintaining strong recall (0.834). This demonstrates that combining controlled undersampling with adversarial oversampling preserves data fidelity while enhancing minority-class representation, leading to improved detection performance under imbalanced conditions.

Table 3: Comparison of class imbalance handling strategies. The best-performing results in the table are shown in bold.

Strategy	Accuracy	Precision	Recall	F1-Score	AUC-PR
Random Undersampling	0.984	0.881	0.801	0.839	0.928
SMOTE	0.986	0.872	0.824	0.847	0.936
ADASYN	0.985	0.865	0.831	0.847	0.934
Proposed Strategy	**0.991**	**0.917**	**0.834**	**0.872**	**0.952**

4.2.3 Adversarial Test Performance (FGSM)

Table 4 presents the model performance under adversarial perturbations generated using the Fast Gradient Sign Method (FGSM). Compared to the clean test results, all models experience noticeable degradation across all metrics, confirming their vulnerability to adversarial attacks. The baseline neural network shows the most severe drop, with recall falling to 0.402 and F1-score to 0.486. Ensemble models like Random Forest and XGBoost show marginal improvement but still lack sufficient robustness. The adversarially trained model shows clear gains over traditional models, reaching an F1-score of 0.597. Notably, the proposed method achieves the best results across all metrics, including the highest accuracy (0.796), precision (0.751), recall (0.569), F1-score (0.648), and AUC-PR (0.693), indicating its effectiveness in mitigating FGSM-based adversarial threats. These results validate the contribution of the integrated adversarial training and DAE-based filtering defense in enhancing robustness. Recent tabular deep learning models such as TabNet and FT-Transformer show improved robustness compared to traditional baselines under FGSM attack; however, they remain less effective than the proposed method, particularly in maintaining higher recall and F1-score under adversarial perturbations.

Table 4: Performance under FGSM attack. The best-performing results in the table are shown in bold.

Model	Accuracy	Precision	Recall	F1-Score	AUC-PR
Baseline NN	0.654	0.612	0.402	0.486	0.591
Random Forest	0.671	0.621	0.438	0.514	0.607
XGBoost	0.689	0.641	0.455	0.527	0.615
TabNet	0.718	0.689	0.498	0.578	0.634
FT-Transformer	0.731	0.701	0.512	0.592	0.645
Adversarial Training Only	0.743	0.702	0.521	0.597	0.642
Proposed Method	**0.796**	**0.751**	**0.569**	**0.648**	**0.693**

4.2.4 Adversarial Test Performance (PGD)

Table 5 reports the model performance under Projected Gradient Descent (PGD) attack, a stronger and more iterative adversarial threat compared to FGSM. As expected, PGD causes further performance degradation across all models. The baseline neural network exhibits the lowest robustness, with a recall of only 0.354 and an F1-score of 0.438. Ensemble models such as Random Forest and XGBoost perform marginally better but remain vulnerable, particularly in recall. The adversarially trained model shows improved resilience with a recall of 0.477 and F1-score of 0.554. The proposed method, integrating both adversarial training and filtering, achieves the highest performance across all evaluated metrics accuracy (0.754), precision (0.721), recall (0.523), F1-score (0.610), and AUC-PR (0.637). These results confirm that the proposed defense is more effective in resisting stronger gradient-based attacks such as PGD, providing a meaningful improvement in robustness without sacrificing detection accuracy. Similarly, TabNet and FT-Transformer demonstrate moderate resilience under the stronger PGD attack compared to traditional baselines; however, their performance remains inferior to the proposed method, highlighting the advantage of the integrated adversarial defense strategy.

Table 5: Performance under PGD attack. The best-performing results in the table are shown in bold.

Model	Accuracy	Precision	Recall	F1-Score	AUC-PR
Baseline NN	0.598	0.582	0.354	0.438	0.541
Random Forest	0.614	0.596	0.387	0.464	0.558
XGBoost	0.633	0.603	0.402	0.474	0.567
TabNet	0.661	0.632	0.438	0.517	0.584
FT-Transformer	0.676	0.648	0.456	0.535	0.596
Adversarial Training Only	0.692	0.655	0.477	0.554	0.593
Proposed Method	**0.754**	**0.721**	**0.523**	**0.610**	**0.637**

4.2.5 Defense Robustness (Delta between Clean and FGSM)

Table 6 summarizes the degradation in key metrics F1-score, recall, and AUC-PR when moving from clean to FGSM-adversarial test conditions. This delta-based evaluation offers a clearer view of each model's vulnerability to perturbation. As expected, the baseline neural network suffers the highest drop across all metrics, with a notable 0.354 decrease in recall and 0.310 in F1-score. Both Random Forest and XGBoost show slightly lower but still significant performance losses. The adversarially trained model reduces this degradation substantially, indicating the benefit of incorporating robustness during training. However, the proposed method demonstrates the smallest performance drop in all metrics: Delta F1 (0.224), Delta Recall (0.265), and Delta AUC-PR (0.259), proving its superior stability and resistance to adversarial distortion. These results reinforce the complementary strength of combining adversarial training with filtering mechanisms.

Table 6: Performance frop (Delta F1, Recall, AUC-PR). The best-performing results in the table are shown in bold.

Model	Delta F1 (↓)	Delta Recall (↓)	Delta AUC-PR (↓)
Baseline NN	0.310	0.354	0.310
Random Forest	0.303	0.333	0.305
XGBoost	0.299	0.327	0.306

(Continued)

Table 6 (continued)

Model	Delta F1 (↓)	Delta Recall (↓)	Delta AUC-PR (↓)
Adversarial Training Only	0.248	0.281	0.294
Proposed Method	**0.224**	**0.265**	**0.259**

4.2.6 Adversarial Detection via Autoencoder

Table 7 presents the performance of the DAE-based filtering mechanism in detecting adversarial samples across different models. The metrics include detection accuracy, false positive rate (FPR), false negative rate (FNR), and the AUC of the reconstruction error curve. The baseline neural network exhibits moderate detection ability, with a detection accuracy of 0.812 and relatively high FNR of 0.314, indicating missed adversarial instances. Ensemble models such as Random Forest and XGBoost show incremental improvements in both detection accuracy and error rates. The adversarially trained model achieves better adversarial detection, with a reduction in both FPR and FNR. The proposed method outperforms all other approaches, achieving the highest detection accuracy (0.878), lowest FPR (0.069), lowest FNR (0.213), and the best reconstruction AUC (0.792). These results confirm that integrating a denoising autoencoder as a pre-filter substantially enhances the model's capacity to identify adversarial inputs without compromising normal operation.

Table 7: Adversarial detection performance using DAE-based filtering. The best-performing results in the table are shown in bold.

Model	Detection Accuracy	FPR	FNR	Reconstruction AUC
Baseline NN	0.812	0.091	0.314	0.721
Random Forest	0.825	0.088	0.298	0.734
XGBoost	0.831	0.082	0.287	0.741
Adversarial Training Only	0.857	0.074	0.244	0.768
Proposed Method	**0.878**	**0.069**	**0.213**	**0.792**

4.2.7 Training and Inference Time

Table 8 compares the computational efficiency of all evaluated models in terms of training time, inference time per sample, and the total number of trainable parameters. As expected, traditional ensemble models such as Random Forest and XGBoost require longer training and inference times due to their complex tree-based structures, with XGBoost reaching a training time of 70.2 s. The baseline neural network is the most lightweight in both parameter count and runtime, but it suffers from lower accuracy and robustness. The adversarially trained model adds a moderate overhead to the baseline NN due to gradient-based augmentation during training. The proposed method, which integrates adversarial training and DAE-based filtering, incurs the highest training (92.6 s) and inference time (0.063 s), along with a modest increase in parameters (0.014 million). However, this added cost is justified by the significant gains in accuracy and robustness across all evaluation scenarios, making the proposed model suitable for deployment in security-critical financial systems where resilience is paramount. In addition to the reported metrics, we further analyze the runtime efficiency of the proposed model. Based on the measured inference time of 0.063 s per sample, the model achieves approximately 15.87 samples per second (FPS), indicating its suitability for near real-time financial fraud detection scenarios. The model also maintains a compact size

of approximately 0.014 million parameters (≈0.056 MB in memory using 32-bit precision), resulting in low memory overhead. This compact representation implies minimal memory requirements across typical CPU and GPU environments, making the approach suitable for deployment in resource-constrained settings. Due to the relatively shallow fully connected architecture, the computational complexity in terms of floating-point operations (FLOPs) remains modest compared to deep convolutional models.

Table 8: Computational efficiency. The best-performing results in the table are shown in bold.

Model	Training Time (s)	Inference Time (s)	Parameters (Millions)
Baseline NN	52.4	0.021	0.012
Random Forest	64.1	0.039	0.056
XGBoost	70.2	0.046	0.084
Adversarial Training Only	85.3	0.059	0.012
Proposed Method	**92.6**	**0.063**	**0.014**

4.2.8 Ablation Study

Table 9 presents the results of an ablation study designed to assess the individual and combined impact of adversarial training and DAE-based filtering under FGSM attack conditions. The baseline configuration, which uses no defense, performs the worst across all metrics. Introducing only adversarial training significantly improves robustness, increasing the F1-score from 0.486 to 0.597 and recall from 0.402 to 0.521. Similarly, using only DAE-based filtering also yields notable improvements, though slightly less effective than adversarial training alone. The combined configuration of our proposed method demonstrates the highest robustness with an F1-score of 0.648, recall of 0.569, and AUC-PR of 0.693. These results clearly show that while each component contributes independently to adversarial robustness, their integration provides complementary benefits and yields the most effective defense. This ablation setup provides a clear breakdown of the individual and combined contributions of adversarial training and DAE-based filtering, allowing us to quantify how each component contributes to overall robustness under adversarial conditions. Quantitatively, adversarial training improves the F1-score from 0.486 to 0.597 (a gain of 0.111), while DAE-based filtering alone increases it to 0.573 (a gain of 0.087). The combined approach further raises the F1-score to 0.648, demonstrating an overall improvement of 0.162 over the baseline and highlighting the complementary contribution of both components.

Table 9: Ablation study under FGSM attack. The best-performing results in the table are shown in bold.

Configuration	Accuracy	Precision	Recall	F1-Score	AUC-PR
Baseline (No Defense)	0.654	0.612	0.402	0.486	0.591
Only Adv. Training	0.743	0.702	0.521	0.597	0.642
Only DAE Filtering	0.721	0.685	0.498	0.573	0.627
Proposed (Both)	**0.796**	**0.751**	**0.569**	**0.648**	**0.693**

These results indicate that adversarial training provides the primary robustness improvement, particularly in enhancing recall and F1-score under attack, while the DAE-based filtering contributes additional complementary gains by identifying and mitigating residual adversarial perturbations.

4.3 Robustness and Adversarial Detection Analysis

Fig. 3 provides a comprehensive visual analysis of model robustness and adversarial detection performance under FGSM attack conditions. Fig. 3a compares F1-scores on clean and adversarial test sets, showing that all models suffer performance degradation when exposed to adversarial perturbations. The baseline neural network exhibits the most severe decline, while ensemble models such as Random Forest and XGBoost demonstrate slightly improved resilience. The adversarially trained model limits this degradation more effectively, whereas the proposed method achieves the smallest reduction in F1-score, maintaining the highest performance under both clean and adversarial settings.

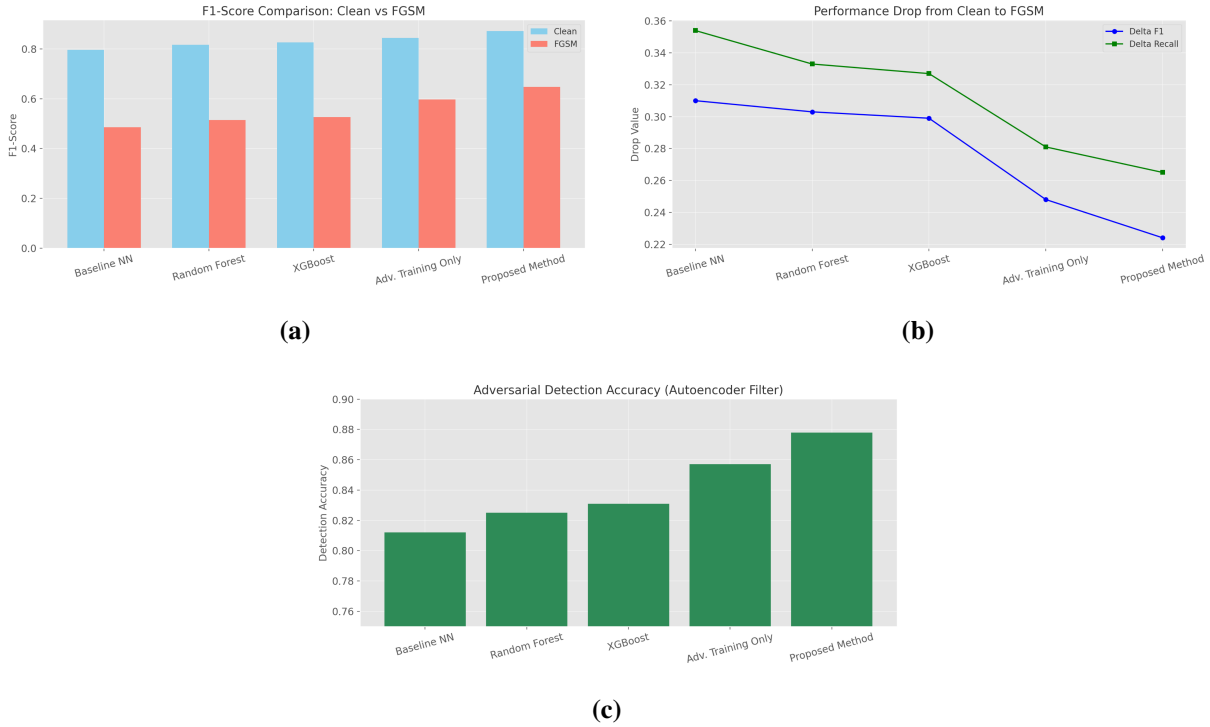

Figure 3: Visual robustness and adversarial detection analysis under FGSM attack. The proposed method demonstrates superior robustness, minimal performance degradation, and the highest adversarial detection accuracy. (**a**) F1-score comparison on clean and FGSM-adversarial test sets. (**b**) Performance degradation measured by Delta F1 and Delta Recall. (**c**) Adversarial detection accuracy using DAE-based filtering.

Fig. 3b illustrates the performance drop from clean to FGSM-adversarial conditions, measured by changes in F1-score and recall. The baseline and ensemble models experience substantial declines, particularly in recall, highlighting their vulnerability to adversarial manipulation. In contrast, adversarial training reduces this degradation, and the proposed method exhibits the lowest delta values across both metrics, confirming its superior stability and robustness. These trends are consistent with the quantitative results reported in Table 6.

Finally, Fig. 3c presents the adversarial detection accuracy achieved using the DAE-based filtering mechanism. While all models benefit from reconstruction-based detection, the baseline neural network shows the weakest separation between clean and adversarial inputs. Ensemble models perform moderately better, and adversarial training further improves detection sensitivity. The proposed method achieves the highest detection accuracy, demonstrating the synergistic benefit of combining adversarial training with

DAE-based filtering. Overall, this visual analysis reinforces the effectiveness of the proposed dual-defense framework in maintaining classification robustness while accurately identifying adversarial inputs.

4.4 Training Dynamics and Convergence Analysis

Fig. 4 illustrates the training dynamics of the proposed model over 20 epochs using loss, accuracy, and F1-score curves for both training and validation sets. Fig. 4a shows a smooth and consistent decrease in both training and validation loss, indicating stable optimization and effective convergence. The close alignment between the two curves suggests that the model generalizes well to unseen data without signs of overfitting, confirming that the chosen regularization strategies and training configuration are appropriate.

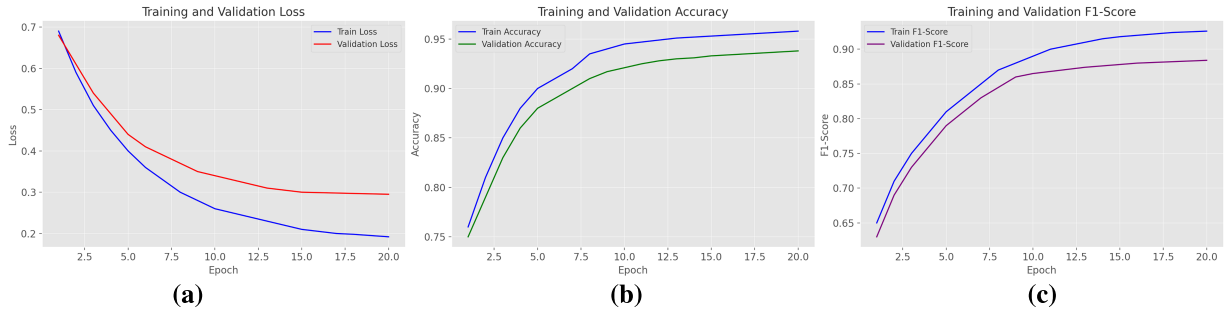

Figure 4: Training dynamics of the proposed model. Consistent loss reduction and steady improvement in accuracy and F1-score indicate stable convergence and strong generalization performance. (**a**) Training and validation loss over epochs. (**b**) Training and validation accuracy over epochs. (**c**) Training and validation F1-score over epochs.

Fig. 4b presents the evolution of training and validation accuracy across epochs. Both curves exhibit a steady upward trend, reflecting continuous improvement in classification performance. The validation accuracy closely follows the training accuracy, indicating stable learning behavior and well-tuned model parameters that avoid overfitting while preserving generalization capability.

Fig. 4c depicts the progression of training and validation F1-scores, which is a critical metric for fraud detection due to extreme class imbalance. The consistent rise in F1-score demonstrates an improving balance between precision and recall as training progresses. The strong alignment between training and validation curves further confirms that the model learns robust and generalizable representations, reinforcing its suitability for real-world financial anomaly detection scenarios.

5 Discussion

The results demonstrate that combining adversarial training with DAE-based filtering improves robustness in credit card fraud detection. This hybrid approach offers a novel contribution to the field of secure AI systems in financial domains, where maintaining both high predictive performance and robustness against adversarial manipulation is critical. Unlike traditional models that often suffer a steep decline in detection accuracy when exposed to adversarial inputs, our method exhibits strong resilience across multiple attack settings, particularly under FGSM and PGD perturbations. This is attributed to the joint use of adversarial training which equips the model to recognize and resist crafted inputs—and the DAE, which acts as a pre-processing filter to detect and isolate anomalous patterns before classification.

From the experimental evaluation, it is evident that the proposed model not only outperforms conventional baselines on clean data but also maintains superior performance under adversarial conditions. In particular, the strong AUC-PR performance highlights the model's effectiveness in handling extreme class

imbalance, as this metric better captures the precision–recall trade-off for rare fraud cases compared to traditional accuracy-based evaluation. The robustness is quantified by a lower performance drop (delta) across F1-score, recall, and AUC-PR, indicating that the model retains its discriminative power even when the input space is maliciously altered. Furthermore, the DAE achieves high adversarial detection accuracy, adding an interpretable and modular defense layer to the pipeline. While the DAE-based filtering mechanism achieves high adversarial detection accuracy, it introduces an inherent trade-off between detecting adversarial samples and potentially flagging difficult but legitimate transactions. In particular, samples that lie near the decision boundary or exhibit atypical but valid fraud patterns may be partially suppressed due to reconstruction smoothing. However, the relatively low false positive and false negative rates observed in the detection results indicate that this trade-off is effectively managed, and the filtering mechanism preserves most meaningful transaction patterns while mitigating adversarial noise. These results imply that the framework is not only effective but also adaptable and extensible to other security-sensitive applications where adversarial threats are prevalent.

Another important consideration in adversarial financial systems is model calibration, which reflects the reliability of predicted confidence scores. In fraud detection applications, well-calibrated probabilities are critical for risk-based decision-making, such as transaction blocking, alert prioritization, or human review. Under adversarial conditions, even when classification accuracy remains stable, confidence estimates may become unreliable, leading to overconfident incorrect predictions. While the proposed framework focuses primarily on improving robustness and detection performance, the integration of adversarial training and filtering mechanisms may also influence calibration behavior. A detailed evaluation of calibration metrics such as expected calibration error (ECE) or Brier score is an important direction for future work to further assess the reliability of predictions in adversarial environments.

From a deployment perspective, the inclusion of the DAE-based filtering mechanism introduces additional inference overhead compared to standard single-stage classifiers. In practical real-time fraud detection pipelines, this additional processing step may increase latency. However, as shown in the computational efficiency analysis, the overall inference time remains low (0.063 s per sample, approximately 15.87 FPS), indicating that the framework is still suitable for near real-time deployment. Moreover, the modular design of the filtering stage allows it to be selectively applied in high-risk scenarios or integrated into asynchronous processing pipelines, thereby mitigating its impact on system latency while preserving robustness against adversarial inputs.

Despite these strengths, several limitations should be acknowledged. First, while the FGSM and PGD attacks provide a solid basis for evaluation, they represent only a subset of the adversarial landscape. More advanced and adaptive attack strategies, including optimization-based attacks such as CW and iterative attacks such as DeepFool, may expose vulnerabilities not captured in this study. In particular, adaptive adversaries that are specifically designed to bypass both adversarial training and DAE-based filtering could reduce the effectiveness of the proposed defense pipeline. Additionally, transferability of adversarial examples across different models and data distributions remains an open concern, as attacks crafted on surrogate models may generalize to the deployed system. Second, although the DAE adds value in detecting adversarial inputs, it introduces computational overhead, which may impact real-time deployment in high-frequency financial systems. Third, although the dataset used in this study is publicly available and widely adopted, the evaluation is limited to a single dataset; therefore, future work will consider cross-dataset validation to further assess the generalization capability of the proposed framework across diverse financial data distributions. Additionally, the hyperparameter selection for both adversarial training and the autoencoder architecture was conducted through empirical tuning, which may not generalize optimally across other

datasets or domains. These factors limit the immediate applicability of the method in resource-constrained or dynamic environments.

Future work should explore several avenues to address these limitations. First, extending the evaluation to include a broader set of attack types and threat models will help assess the generalizability of the defense. Incorporating certified defenses or randomized smoothing may offer theoretical robustness guarantees that complement empirical resilience. Second, adaptive DAE architectures, potentially driven by neural architecture search or meta-learning, could reduce latency while maintaining high detection accuracy. Finally, integrating explainability mechanisms such as saliency maps or reconstruction error visualization can provide stakeholders with better insights into why inputs are flagged as adversarial, enhancing the trust and transparency of AI systems in financial operations.

In summary, this study introduces and validates a novel dual-defense strategy for adversarial robustness in credit card fraud detection. The combination of adversarial training and DAE-based filtering yields strong empirical results under attack, and lays the foundation for future developments in secure, interpretable, and efficient AI-driven financial systems.

6 Conclusions

This study proposes a dual-defense framework combining adversarial training and DAE-based filtering to improve the robustness of AI models against adversarial attacks in credit card fraud detection. The method outperforms baseline models on clean data and maintains superior resilience under FGSM and PGD attacks, achieving the highest scores in F1, recall, and AUC-PR metrics. The DAE component enhances detection accuracy by identifying perturbed inputs through reconstruction error, while adversarial training equips the classifier to better withstand adversarial perturbations. The approach is empirically validated through extensive experiments, including ablation studies and quantitative evaluations across multiple metrics. While the model introduces additional computational cost and may require tuning for deployment in other settings, its modular design and consistent performance suggest strong potential for broader application in security-critical financial systems. Future work may explore more complex attack models, integrate certified defenses, and adapt the framework for real-time environments or other high-risk domains.

Acknowledgement: We would like to express our sincere gratitude to the Advanced Machine Intelligence Research Lab (AMIR Lab) for providing valuable supervision and necessary resources that significantly supported the successful completion of this review work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and study design: Mohammed Saad Javeed; Data collection: Jannatul Maua; Methodology and implementation: Mohammed Saad Javeed, Hashibul Ahsan Shoaib; Formal analysis and validation: Muhammad Firoz Mridha; Writing—original draft preparation: Mohammed Saad Javeed, Jannatul Maua; Writing—review and editing: Jungpil Shin, Taro Suzuki; Supervision: Jungpil Shin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All data analyzed in this study are publicly available from the cited sources.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as Editorial Board Member of this journal, Jungpil Shin had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Hafez IY, Hafez AY, Saleh A, Abd El-Mageed AA, Abohany AA. A systematic review of AI-enhanced techniques in credit card fraud detection. J Big Data. 2025;12(1):6. doi:10.1186/s40537-024-01048-8.
2. Wang Z, Shen Q, Bi S, Fu C. AI empowers data mining models for financial fraud detection and prevention systems. Procedia Comput Sci. 2024;243:891–9.
3. Chen Y, Zhao C, Xu Y, Nie C, Zhang Y. Deep learning in financial fraud detection: innovations, challenges, and applications. Data Sci Manag. 2026;9(2):100162. doi:10.1016/j.dsm.2025.08.002.
4. Mohamed N. Cutting-edge advances in AI and ML for cybersecurity: a comprehensive review of emerging trends and future directions. Cogent Bus Manag. 2025;12:2518496.
5. Narender M, Anand AJ. Artificial intelligence in financial fraud detection. In: Handbook of AI-driven threat detection and prevention. Boca Raton, FL, USA: CRC Press; 2025. p. 193–207.
6. Aldahdooh A, Hamidouche W, Fezza SA, Déforges O. Adversarial example detection for DNN models: a review and experimental comparison. Artif Intell Rev. 2022;55(6):4403–62.
7. Pelekis S, Koutroubas T, Blika A, Berdelis A, Karakolis E, Ntanos C, et al. Adversarial machine learning: a review of methods, tools, and critical industry sectors. Artif Intell Rev. 2025;58(8):226. doi:10.1007/s10462-025-11147-4.
8. Lunghi D, Simitsis A, Caelen O, Bontempi G. Adversarial learning in real-world fraud detection: challenges and perspectives. In: Proceedings of the Second ACM Data Economy Workshop; 2023 Jun 18; Seattle, WA, USA. p. 27–33.
9. Wawrowski Ł, Biczyk P, Ślęzak D, Sikora M. Adversarial attacks detection method for tabular data. Mach Learn Knowl Extr. 2025;7(4):112.
10. Chang V, Ali B, Golightly L, Ganatra MA, Mohamed M. Investigating credit card payment fraud with detection methods using advanced machine learning. Information. 2024;15(8):478. doi:10.3390/info15080478.
11. Talukder MA, Khalid M, Uddin MA. An integrated multistage ensemble machine learning model for fraudulent transaction detection. J Big Data. 2024;11(1):168. doi:10.1186/s40537-024-00996-5.
12. Al-Daoud KI, Abu-AlSondos IA. Robust AI for financial fraud detection in the GCC: a hybrid framework for imbalance, drift, and adversarial threats. J Theor Appl Electron Commer Res. 2025;20(2):121. doi:10.3390/jtaer20020121.
13. Karthika J, Senthilselvi A. Smart credit card fraud detection system based on dilated convolutional neural network with sampling technique. Multimed Tools Appl. 2023;82(20):31691–708.
14. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. Proc AAAI Conf Artif Intell. 2021;35(8):6679–87. doi:10.1609/aaai.v35i8.16826.
15. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Adv Neural Inf Process Syst. 2021;34:18932–43. doi:10.48550/arxiv.2106.11959.
16. Villegas-Ch W, Jaramillo-Alcázar A, Luján-Mora S. Evaluating the robustness of deep learning models against adversarial attacks: an analysis with FGSM, PGD and CW. Big Data Cogn Comput. 2024;8(1):8.
17. Naseem ML. Trans-IFFT-FGSM: a novel fast gradient sign method for adversarial attacks. Multimed Tools Appl. 2024;83(29):72279–99.
18. Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M. Adversarial machine learning: a taxonomy and terminology of attacks and mitigations. Gaithersburg, MD, USA: National Institute of Standards and Technology; 2025.
19. Gangwani P, Soni J, Almonte A, Perez-Pons A, Upadhyay H. Adversarial AI training to mitigate cyber attacks. In: Proceedings of the 2025 IEEE International Conference on Artificial Intelligence Testing (AITest); 2025 Jul 21–24; Tucson, AZ, USA. New York, NY, USA: IEEE; 2025. p. 54–61.
20. Al Roken N, Hacid H, Bouridane A, Hussian A. Hybrid adversarial retraining approach for a proactive network intrusion detection system. IEEE Open J Commun Soc. 2026;7:5350–64. doi:10.1109/ojcoms.2026.3694298.
21. Zavvar M, Jafari M, Pour NM, Kiaei AA, Zavvar MH, Heidari A, et al. A hybrid deep learning framework using synthetic oversampling, autoencoder, convolutional neural networks, and an attention mechanism for credit card fraud detection. J Big Data. 2026;13(1):21. doi:10.1186/s40537-025-01331-2.
22. Heidari A, Jafari N, Zeadally S. Designing efficient anomaly detection systems using deep learning techniques. Int J Pervasive Comput Commun. 2026;22(1–2):31–55. doi:10.1108/ijpcc-08-2024-0262.