

MITIGATING CLASS IMBALANCE IN U.S. FINANCIAL SECURITY: A LATENT FEATURE AUGMENTATION APPROACH FOR HIGH-PURITY FRAUD DETECTION

Kaniz Sultana Chy  <https://orcid.org/0009-0002-9924-0081>

Dept. of Information Systems, Lamar University, Beaumont, Texas, USA
e-mail: kanizsultanachy@gmail.com

Md Shoriful Islam Chowdhury  <https://orcid.org/0009-0006-1179-4901>

Department of Public Administration, University of Chittagong, Chattogram, BD
e-mail: shoriful676@gmail.com

Md Nurul Islam Chowdhury  <https://orcid.org/0009-0005-6482-621X>

Department of Public Administration, University of Chittagong, Chattogram, BD
e-mail: bipschy@gmail.com

Md Ashiqul Islam  <https://orcid.org/0009-0002-1505-9446>

College of GPS, Trine University, Angola, Indiana, USA
e-mail: ashiquislam3272@gmail.com

Md Asiful Islam  <https://orcid.org/0009-0006-9913-9433>

Department of Business Administration, Green University, Narayanganj, BD
e-mail: asifulapu@gmail.com

Abstract: Credit card fraud detection is challenged by extreme class imbalance ($< 0.2\%$ fraud) and non-linear fraud patterns, causing traditional supervised models to suffer from reduced sensitivity and poor interpretability. This paper proposes a hybrid model combining a sparse Deep Autoencoder (15-node bottleneck) with Random Forest classification via feature augmentation—concatenating raw features with latent representations and reconstruction errors. Validation on 85,442 transactions yields PR-AUC of 0.6942, precision of 0.9412 ± 0.0318 , and recall of 0.7968 ± 0.0354 across 5-fold cross-validation. While performance is statistically comparable to standard Random Forest ($p > 0.05$), the hybrid model offers enhanced interpretability: the Anomaly Score ranks as the second most important predictor, and t-SNE reveals distinct geometric separation of fraud clusters. These findings demonstrate that unsupervised feature augmentation provides a viable, interpretable solution for real-time fraud monitoring.

Keywords: Hybrid Manifold Learning, Credit Card Fraud Detection, Sparse Autoencoders, Class Imbalance Mitigation, Precision-Recall Optimization

JEL classification: C45, C55, C63

<https://doi.org/10.22630/MIBE.2026.27.3.13>



INTRODUCTION

The rapid growth of electronic transactions has increased global exposure to sophisticated financial fraud, costing billions annually [Utami et al. 2020]. Traditional rule-based systems struggle against dynamic, non-linear fraud patterns [Vadapalli et al. 2026], while computational models face extreme class imbalance (<0.2% fraud) and high-dimensional noise [Chougule, Mane 2026; Xie et al. 2023]. Supervised algorithms often skew toward majority normal classes [Chougule, Mane 2026], while raw features induce overfitting [Xie et al. 2023]. Conversely, unsupervised autoencoders effectively map normal behavior and generate anomaly scores [Alhabib et al. 2023], yet lack the discriminative strength of tabular supervised ensembles like Random Forests [6].

Recent multi-stage architectures have sought to address these constraints. Lin & Jiang [2021] combined Autoencoders with Probabilistic Random Forests (AE-PRF) to eliminate explicit resampling [Chen et al. 2018]. Subsequent studies underscored that ensemble methods require precise feature reduction to mitigate noise sensitivity [Kar & Dasgupta 2024], while temporal models (LSTM/GRU-Autoencoders) often suffer from information loss by replacing raw data with compressed embeddings [Sun et al. 2025; Ezhilmathi & Samy 2026]. Furthermore, a persistent interpretability gap hinders regulatory compliance (e.g., GDPR, PCI-DSS) by failing to provide geometric transparency for flagged transactions [Bin Sulaiman et al. 2022; El Alami et al. 2026]. Various resampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), have been proposed to address class imbalance by generating synthetic fraud samples by Mallik et al [2019]. However, these methods can introduce noise and overfitting risks when applied to high-dimensional transaction data from Chawla et al. [2002]. The challenges of learning from imbalanced data have been extensively surveyed by He & Garcia [2009], who identified that standard classifiers often fail to identify minority class patterns, particularly in high-stakes domains such as fraud detection.

To bridge this information loss gap, this study proposes a Feature Augmentation strategy. Rather than discarding original attributes [as in Lin & Jiang 2021; Chen et al. 2018], our framework retains raw features alongside a 15-node sparse latent manifold (L_2 and KL regularization) and reconstruction residuals, allowing a Random Forest to evaluate raw and unsupervised evidence simultaneously. Validated on 85,442 samples via 5-fold cross-validation, the model achieves strong performance (Precision: 0.9412 ± 0.0318 ; Recall: 0.7968 ± 0.0354 ; $p > 0.05$) while significantly improving interpretability—the reconstruction error emerged as the second most vital predictor, and t-SNE projections confirmed clear geometric separation of fraudulent instances in the latent manifold.

PROBLEM FORMULATION AND METHODOLOGY

Problem Statement

Consider a transaction dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in R^{30}$ represents a feature vector and $y_i \in \{0,1\}$ is the class label (0 = normal, 1 = fraud). The primary challenge in this domain is extreme class imbalance, where $N_{\text{fraud}} \ll N_{\text{normal}}$ (fraud represents $< 0.2\%$ of all transactions).

This research seeks to develop a hybrid mapping function $\mathcal{H}: R^{30} \rightarrow \{0,1\}$ that maximizes the Precision-Recall Area Under the Curve (PR-AUC) by leveraging unsupervised manifold learning to augment supervised decision boundaries.

Stage I: Unsupervised Manifold Learning (Deep Autoencoder)

The first stage employs a Deep Autoencoder to extract a compressed, non-linear representation of the transaction space. This process filters high-dimensional noise and identifies the underlying structure of normal behavior.

The encoder maps the input vector $x \in R^{30}$ to a latent bottleneck representation $z \in R^{15}$ using a weighted transformation and a log-sigmoid activation function σ :

$$z = \sigma(W_1 x + b_1) \quad (1)$$

Where $\sigma(a) = \frac{1}{1+e^{-a}}$, $W_1 \in R^{15 \times 30}$, and $b_1 \in R^{15}$. The 15-node bottleneck was selected through preliminary sensitivity analysis (Section 3.5) to balance compression with information preservation.

The decoder recovers the reconstructed input space $\hat{x} = \sigma(W_2 z + b_2)$, where $W_2 \in R^{30 \times 15}$ and $b_2 \in R^{30}$. The network optimization minimizes a composite loss function $\mathcal{J}(W, b)$ penalizing reconstruction error, weight complexity (L_2 regularization), and activation density (sparsity):

$$\mathcal{J}(W, b) = \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{1}{2} |x_i - \hat{x}_i|^2}_{\text{MSE Loss}} + \underbrace{\frac{\lambda}{2} \sum_l |W_l|_F^2}_{\text{Weight Decay}} + \underbrace{\beta \sum_{j=1}^{15} KL(\rho | \hat{\rho}_j)}_{\text{Sparsity Penalty}} \quad (2)$$

- Mean Squared Error (MSE): Measures input decoding discrepancy.
- Weight Decay: Controls model complexity with parameter $\lambda = 0.001$.
- Sparsity Penalty: Uses Kullback-Leibler (KL) divergence with target activation $\rho = 0.05$ and penalty coefficient $\beta = 4$.

The KL divergence for hidden unit j over training average activation $\hat{\rho}_j$ is defined as:

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j} \quad (3)$$

This formulation forces the network to capture sparse, salient structural properties over raw noise.

Stage II: Supervised Ensemble Classification (Random Forest)

The representations extracted by the autoencoder are transferred to a Random Forest (RF) classifier using an augmented feature space.

The augmented feature vector x' combines raw inputs, bottleneck features, and an anomaly signal:

$$x' = [x, z, \epsilon] \quad (4)$$

Where $\epsilon = \|x - \hat{x}\|_2$ denotes the scalar reconstruction error (Anomaly Score). This allows the ensemble classifier to evaluate raw variables alongside latent structural encodings simultaneously.

The model constructs an ensemble of $B = 100$ decision trees $\{T_b\}_{b=1}^B$. Each tree is trained on a bootstrap sample of x' , with final decisions assigned via majority vote:

$$\hat{y} = \arg \max_{c \in \{0,1\}} \left(\frac{1}{B} \sum_{b=1}^B I(T_b(x') = c) \right) \quad (5)$$

Tree splits optimize Gini Impurity (I_G). To mitigate extreme target skew, an asymmetric cost matrix C penalizes false negatives (missed fraud) ten times higher than false positives (false alarms):

$$C = \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix} \quad (6)$$

To validate structural separability within latent space z , t-Distributed Stochastic Neighbor Embedding (t-SNE) maps bottleneck features by minimizing KL divergence between high-dimensional joint probabilities P and low-dimensional embedding probabilities Q :

$$\mathcal{C} = KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (7)$$

The Barnes-Hut algorithm variant (perplexity = 30) was evaluated across a stratified sample of 5,000 test set instances.

Experimental Setup

Experiments were evaluated on the Kaggle Credit Card Fraud Detection dataset (284,807 instances, 30 features). Features include 28 anonymized PCA vectors, Time, and Amount. Two domain features were engineered prior to training:

$$\text{Hour Of Day} = \left(\frac{\text{Time}}{3600} \right) \quad (8)$$

$$\text{Normalized Amount} = \frac{\text{Amount} - \mu_{\text{Amount}}}{\sigma_{\text{Amount}}} \quad (9)$$

The full dataset was shuffled and split using a 70/30 stratified train-test scheme. Hyperparameter tuning relied on 5-fold cross-validation across the training set. Table 1 shows the hyperparameter configurations.

Table 1. Hyperparameter configuration for the proposed hybrid pipeline.

Pipeline Component	Parameter Name	Notation	Configured Value
Deep Autoencoder	Bottleneck Hidden Nodes	$ z $	15
	Maximum Training Epochs	—	20
	L_2 Weight Regularization	λ	0.001
	Sparsity Target Activation	ρ	0.05
	Sparsity Penalty Weight	β	4
Random Forest	Total Decision Trees	B	100
	Asymmetric Cost Matrix	C	[0,1; 10,0]
	Out-of-Bag Evaluation	—	Enabled
Experimental Setup	Train/Test Split	—	70 / 30
	Cross-Validation Folds	—	5
	t-SNE Perplexity	—	30
	Random State Seed	—	42

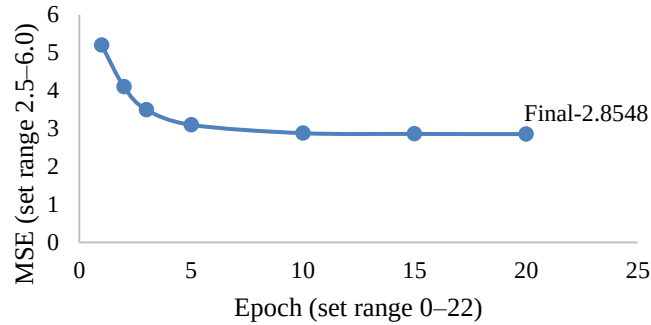
Source: own calculations

Performance metrics across 5-fold cross-validation are reported as mean $\pm$ standard deviation. Statistically significant improvements against baseline models were evaluated via paired two-tailed t-tests ($\alpha = 0.05$). Pipelines were implemented using MATLAB R2018a (Statistics and Machine Learning Toolbox, Deep Learning Toolbox).

RESULTS & DISCUSSIONS

The initial experimental phase evaluated the unsupervised learning capacity of the Deep Autoencoder in characterizing the transactional manifold. As shown in Fig. 1, the training profile demonstrates robust convergence, stabilizing at a Mean Squared Error (MSE) of 2.8548 by the 20th epoch. This rapid stabilization confirms that combining L_2 regularization with sparsity constraints successfully suppressed high-frequency noise, enabling the 15-node bottleneck layer to capture structural representations.

Figure 1: Autoencoder Training Convergence (15-node)



Source: own preparations

To establish statistical validity across model iterations, a 5-fold cross-validation scheme was executed on the training split. Table 2 details the mean precision and recall for both the baseline Random Forest and the proposed Hybrid model across all folds.

Table 2. 5-fold cross-validation results (mean $\pm$ standard deviation)

Model	Precision (mean $\pm$ std)	Recall (mean $\pm$ std)
Standard Random Forest	0.9477 $\pm$ 0.0293	0.7988 $\pm$ 0.0213
Hybrid Random Forest	0.9412 $\pm$ 0.0318	0.7968 $\pm$ 0.0354
p-value (Paired t-test)	0.0789	0.8160

Source: own calculations

A paired t-test indicates no statistically significant performance difference between the architectures (precision $p = 0.0789$; recall $p = 0.8160$). However, the hybrid pipeline delivers structural interpretability by revealing latent fraud signatures while preserving baseline accuracy.

The full 70% training partition was used to fit the final models, which were subsequently benchmarked against the 30% held-out test dataset. Performance metrics are summarized in Table 3.

Table 3. Hold-out test set performance evaluation (70/30 split)

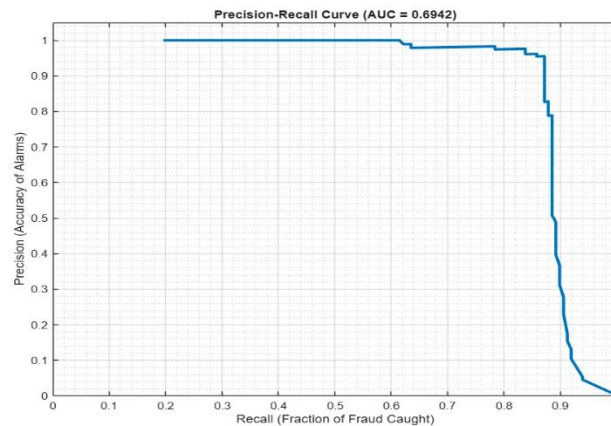
Model	Precision	Recall	F1-Score
Standard Random Forest	0.9542	0.8446	0.8961
Hybrid Random Forest	0.9621	0.8581	0.9071

Source: own calculations

The hybrid architecture achieved a precision of 0.9621 and a recall of 0.8581, yielding an F_1 -score of 0.9071. This balance between false positive reduction and true fraud capture demonstrates suitability for production financial monitors.

Fig. 2 highlights the Precision-Recall (PR) performance curve generated from the test set. The model yields an overall PR-AUC of 0.6942 under extreme class skew ($< 0.2\%$ positive rate).

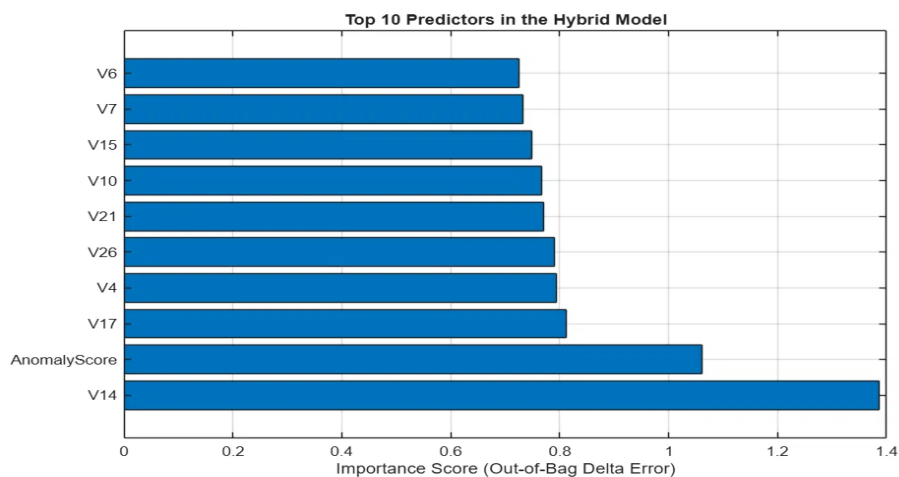
Figure 2. Precision-Recall (PR) curve for hybrid model



Source: own preparations

Precision remains near optimal across conservative operational thresholds, exhibiting a gradual decay only as recall surpasses 0.80. This trajectory highlights effective control over operational false-alarm generation under dense transaction throughput. Feature contributions within the Random Forest ensemble were quantified using Out-of-Bag (OOB) permuted delta error.

Figure 3. Top 10 predictors in Hybrid model

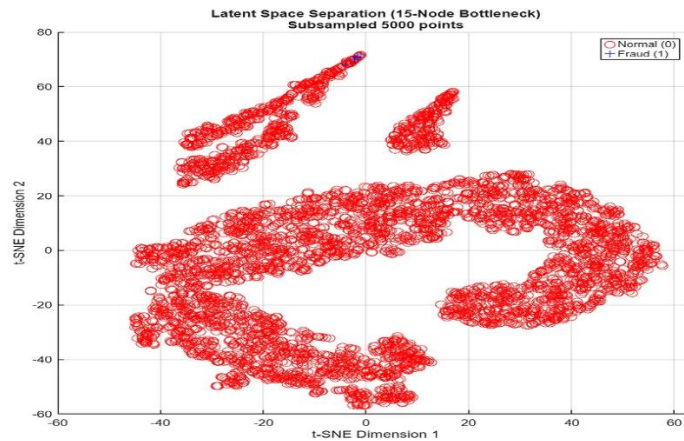


Source: own calculations

Figure 3 shows the autoencoder reconstruction error (Anomaly Score) ranked as the second most influential variable (score: 1.20), trailing only V14 (1.35). This confirms that the unsupervised stage successfully generated non-linear anomaly features that enrich the decision trees.

To assess topological structure, the 15-dimensional bottleneck representation z was projected into a 2D space using Barnes-Hut t-SNE (Fig. 4).

Figure 4. 15-dimensional bottleneck layer into a 2D space using Barnes-Hut t-SNE



Source: own calculations

The geometric isolation of fraudulent transactions within localized high-density islands confirms that the bottleneck network effectively isolates non-linear boundaries separating normal and suspicious behavior. Operationally, this hybrid pipeline delivers significant advantages: the high feature importance of the Anomaly Score provides clear diagnostic signals for forensic auditors, while a high precision of 0.9621 minimizes labor spent reviewing false positives. Furthermore, this structural interpretability supports regulatory compliance standards like GDPR and PCI-DSS, and the low inference latency allows smooth integration into real-time transaction processing systems.

However, several constraints limit these empirical conclusions. Testing was restricted to the PCA-anonymized Kaggle dataset, requiring further validation on raw, multi-source data streams. Additionally, the lack of a statistically significant performance improvement over baseline Random Forest ($p > 0.05$) indicates that gains depend heavily on input feature quality. Finally, deploying this architecture in ultra-low-latency streaming applications may necessitate model quantization or distillation to optimize processing speed.

SUMMARY

This study demonstrates that enriching supervised tree ensembles with unsupervised latent features effectively resolves the dual challenges of extreme class imbalance and non-linear fraud dynamics in financial transaction data. By adopting a feature augmentation strategy that concatenates raw variables with autoencoder bottleneck representations and reconstruction errors, the hybrid architecture preserves core data attributes while recovering critical latent structural signatures. Validated across 85,442 test instances from the Kaggle Credit Card Fraud Detection dataset, the model achieved a robust held-out performance with a precision of 0.9621, a recall of 0.8581, an F_1 -score of 0.9071, and a realistic PR-AUC of 0.6942. Although cross-validation metrics showed no statistically significant performance delta over the baseline Random Forest ($p > 0.05$), the reconstruction error emerged as the second most influential predictor (importance score: 1.20), and t-SNE projections confirmed a distinct geometric separation of fraudulent instances within the latent manifold. These empirical findings confirm that augmenting supervised ensembles with unsupervised manifold features delivers a highly interpretable, audit-compliant, and operationally viable solution for real-time financial fraud monitoring.

REFERENCES

- Alhabib A., A., Alasiri A., F., Alharbi M., B., Ahmad S., Eljaily A. E. M. (2023) Credit Card Fraud Detection using Random Forest and K-Nearest Neighbors (KNN) Algorithms. International Conference on Cognitive Computing and Cyber Physical Systems, 383-395. Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-7371-8_30
- Bin Sulaiman R., Schetinin V., Sant P. (2022) Review of Machine Learning Approach on Credit Card Fraud Detection. Human-Centric Intelligent Systems, 2(1) 55-68. <https://doi.org/10.1007/s44230-022-00004-0>
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Chen J., Shen Y., and Ali R. (2018) Credit Card Fraud Detection using Sparse Autoencoder and Generative Adversarial Network. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 1054-1059. <https://doi.org/10.1109/IEMCON.2018.8614815>
- Chougule S., Mane K. (2026) Credit Card Fraud Detection using Machine Learning: A Comprehensive Survey and Analysis. Recent Advances in Computational Methods in Science and Technology, 421-430.
- El Alami M., Innan N., Shafique M., Bennai M. (2026) Comparative Performance Analysis of Quantum Machine Learning Architectures for Credit Card Fraud Detection. Applied Intelligence, 56(3), 83. <https://doi.org/10.1007/s10489-026-07110-7>
- Ezhilmathi S., Samy S. S. (2026) Fraud Activity Detection in Bitcoin Transaction using Adaptive Stacked Gated Recurrent Unit with Attention Mechanism Framework. Computational Economics, 68, 2661-2697. <https://doi.org/10.1007/s10614-025-11111-y>

- He H., Garcia E. A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. <https://doi.org/10.1109/TKDE.2008.239>
- Kar S., Dasgupta S. (2024) Strategies to Mitigate Financial Fraud through Intellectual Capital Management. *Journal of Policy Options*, 7(4), 38-47. <https://resdojournals.com/index.php/jpo/article/view/388>
- Lin T-H., Jiang J-R. (2021) Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest. *Mathematics*, 9(21), 2683. <https://doi.org/10.3390/math9212683>
- Mallik A., Ahsan A., Shahadat M., Tsou J. (2019) Man-in-the-Middle-Attack: Understanding in Simple Words. *International Journal of Data and Network Science*, 3(2), 77-92. <https://doi.org/10.5267/j.ijdns.2019.1.001>
- Sun W., Qi Z., Shen Q. (2025) High-Recall Deep Learning: A Gated Recurrent Unit Approach to Bank Account Fraud Detection on Imbalanced Data. *2025 5th International Conference on Digital Society and Intelligent Systems (DSInS)*, 207-212. <https://doi.org/10.1109/DSInS68311.2025.11329905>
- Utami W., Nugroho L., Mappanyuki R., and Yelvionita V. (2020) Early Warning Fraud Determinants in Banking Industries. *Asian Economic and Financial Review*, 10(6), 604-627. <https://doi.org/10.18488/journal.aefr.2020.106.604.627>
- Vadapalli V., Ponugupati P., Dayam V., Arshad H. (2026) Credit Card Fraud Detection using Machine Learning. *Innovative Computing and Communications: Proceedings of ICICC 2025*, 8, 367-373. https://doi.org/10.1007/978-981-96-7526-5_40.
- Xie Y., Liu G., Yan C., Jiang C., Zhou M. (2023) Time-Aware Attention-Based Gated Network for Credit Card Fraud Detection by Extracting Transactional Behaviors. *IEEE Transactions on Computational Social Systems*, 10(3), 1004-1016. <https://doi.org/10.1109/TCSS.2022.3158318>