Expert Systems With Applications

Ensemble Machine Learning with XGBoost and Deep Neural Networks for Imbalanced Fraud Detection of Financial Transactions

--Manuscript Draft--

Manuscript Number:	ESWA-D-25-29762
Article Type:	Full length article
Section/Category:	3.2 Data analytics and mining
Keywords:	Fraud detection; Machine learning; XGBoost; Deep learning; Ensemble models
Corresponding Author:	Shengkun Xie, PhD Toronto Metropolitan University Ted Rogers School of Management Toronto, ON CANADA
First Author:	Shengkun Xie, PhD
Order of Authors:	Shengkun Xie, PhD
	Nguyen Duy Anh Luong
Abstract:	The rapid digitalization of financial services has transformed e-commerce, mobile payments, and on-line banking but has also increased exposure to fraudulent transaction activities. Effective fraud detection of financial transactions requires decision-support systems that can identify abnormal patterns in large-scale, high-dimensional, and highly imbalanced transaction data. This study proposes a hybrid ensemble approach that integrates Enhanced XGBoost with a Deep Neural Network to detect credit card fraud. Seven supervised machine learning models were systematically evaluated under severe class imbalance using stratified validation and multi-metric assessment including Recall, Precision, F1, ROC-AUC,and PR-AUC. The Enhanced XGBoost model achieved a ROC-AUC of 0.968 and F1-score of 0.72, while the proposed XGBoost + DNN ensemble delivered the best overall performance with ROC-AUC = 0.973, F1-score = 0.74, and PR-AUC = 0.7897, outperforming Random Forest (ROC-AUC = 0.941) and standard DNNs (ROC-AUC = 0.961). SHAP-based feature attribution revealed that transaction amount, cardholder behavior, and temporal patterns were the most discriminative predictors. The results demonstrate that the proposed ensemble offers a scalable, accurate, and interpretable solution for real-world fraudulent detection.

Cover Letter

Dear Editor,

I am pleased to submit our manuscript entitled *"Ensemble Machine Learning with XGBoost and Deep Neural Networks for Imbalanced Fraud Detection of Financial Transactions"* for consideration in *Expert Systems with Applications*.

The rapid digitalization of financial services has created unprecedented opportunities but also heightened exposure to fraudulent activities. Traditional rule-based approaches are increasingly ineffective against evolving fraud tactics, motivating the need for scalable and interpretable machine learning systems. This study contributes to that goal through a hybrid ensemble framework that integrates Enhanced XGBoost with a Deep Neural Network (DNN) to address extreme class imbalance, high-dimensional transaction features, and the interpretability challenges that limit real-world deployment.

The novelty of this work lies in three aspects:

1. **A hybrid ensemble model** that synergizes the structured learning capability of XGBoost with the representational strength of DNNs, achieving superior detection performance under severe imbalance.
2. **An adaptive evaluation framework** combining class weighting, focal loss, and threshold optimization to improve both recall and precision for minority-class detection.
3. **SHAP-based interpretability analysis**, offering global and local explanations of feature influence, thereby enhancing transparency and regulatory compliance.

Empirical results demonstrate that the proposed hybrid ensemble achieves state-of-the-art performance on the IEEE-CIS fraud detection dataset, delivering both scalability and explainability—qualities aligned with *Expert Systems with Applications*' emphasis on intelligent, practical solutions for real-world problems.

We confirm that this manuscript is original, has not been published elsewhere, and is not under consideration by any other journal. We believe it will be of strong interest to the journal's readership in the fields of machine learning, financial analytics, and applied artificial intelligence.

Thank you for considering our submission. We look forward to your evaluation and feedback.

Sincerely,

Shengkun Xie

Highlights

Ensemble Machine Learning with XGBoost and Deep Neural Networks for Imbalanced Fraud Detection of Financial Transactions

Nguyen Duy Anh Luong,Shengkun Xie*

- Hybrid ensemble of Enhanced XGBoost and DNN achieves state-of-the-art results.
- Imbalance-aware Class-weighting and SHAP improve recall and ensure explainability.
- Systematic evaluation framework validates robustness under severe class imbalance.

Ensemble Machine Learning with XGBoost and Deep Neural Networks for Imbalanced Fraud Detection of Financial Transactions

Nguyen Duy Anh Luong, Shengkun Xie*

Global Management Studies, Ted Rogers School of Management, Toronto Metropolitan University, 350 Victoria Street, Toronto, ON M5B 2K3,Canada

ARTICLE INFO

Keywords:
Fraud detection
Machine learning
XGBoost
Deep learning
Ensemble models

ABSTRACT

The rapid digitalization of financial services has transformed e-commerce, mobile payments, and on-line banking but has also increased exposure to fraudulent transaction activities. Effective fraud detection of financial transactions requires decision-support systems that can identify abnormal patterns in large-scale, high-dimensional, and highly imbalanced transaction data. This study proposes a hybrid ensemble approach that integrates Enhanced XGBoost with a Deep Neural Network to detect credit card fraud. Seven supervised machine learning models were systematically evaluated under severe class imbalance using stratified validation and multi-metric assessment including Recall, Precision, F1, ROC-AUC,and PR-AUC. The Enhanced XGBoost model achieved a ROC-AUC of 0.968 and F1-score of 0.72, while the proposed XGBoost + DNN ensemble delivered the best overall performance with ROC-AUC = 0.973, F1-score = 0.74, and PR-AUC = 0.7897, outperforming Random Forest (ROC-AUC = 0.941) and standard DNNs (ROC-AUC = 0.961). SHAP-based feature attribution revealed that transaction amount, cardholder behavior, and temporal patterns were the most discriminative predictors. The results demonstrate that the proposed ensemble offers a scalable, accurate, and interpretable solution for real-world fraudulent detection.

1. Introduction

The rapid digitalization of financial services (Broby, 2021) has fundamentally transformed the way individuals and businesses conduct transactions across e-commerce, mobile payments, and on-line banking. While these innovations have enabled unprecedented convenience and efficiency, they have also introduced new vulnerabilities, exposing financial systems to increasingly sophisticated fraudulent activities. Credit card fraud, in particular, remains one of the most pervasive and costly financial crimes, causing direct monetary losses and eroding public trust in financial institutions (Ding, Ruan, Wang, & Liu, 2025). Accurately identifying fraud is challenging because malicious activities are often concealed within massive streams of legitimate transactions, making traditional detection approaches insufficient.

Early fraud detection systems were predominantly rule-based, relying on static thresholds and handcrafted anomaly detection rules to flag suspicious transactions (Bolton & Hand, 2002). Although such approaches were initially effective, they fail to adapt to evolving fraudulent patterns. Fraudsters continuously modify their tactics, rendering fixed rules obsolete and leading to high false-positive rates while allowing genuine fraud cases to go undetected (Ding et al., 2025). These limitations have driven a paradigm shift toward machine learning (ML) and deep learning (DL), which can automatically learn complex, non-linear patterns from data, adapt to emerging fraud tactics, and uncover relationships that are difficult to model manually (Roseline, Naidu, Pandi, alias Rajasree, & Mageswari, 2022).

ML has shown particular strength in analyzing large-scale, high-dimensional transaction data (Talukder, Khalid, & Uddin, 2024; Zhang et al., 2019). Cost-sensitive learning has been proposed to explicitly account for the financial consequences of misclassification, striking a balance between minimizing false positives and maximizing fraud detection (Roseline et al., 2022). To address scalability requirements, distributed learning frameworks have been developed to process high-velocity data streams in near real time (Theodorakopoulos, Theodoropoulou, Tsimakis, & Halkiopoulos, 2025). In addition, various algorithms have been explored: support vector machines and neural networks effectively capture non-linear relationships, whereas gradient boosting methods such as XGBoost offer robust performance and strong predictive accuracy for tabular, structured data (Noviandy et al., 2023).

Despite these advances, fraudulent detection still faces several challenges. One of the most persistent is the extreme class imbalance between legitimate and fraudulent transactions, where fraud often accounts for less than 1% of all records (Breskuvienė & Dzemyda, 2024). This imbalance biases models toward the majority class and reduces recall for fraudulent cases, which is particularly costly in real-world applications. Existing solutions include active learning to prioritize informative samples (Riskiyadi, 2024), ensemble boosting methods tailored to minority classes, and resampling strategies such as SMOTE and undersampling to rebalance the data distribution (Ludera, 2021).

Another major challenge is the high dimensionality and structural complexity of financial transaction data, which involve non-linear feature interactions. Feature engineering has therefore become central to improving model performance, with techniques such as time-window aggregation to capture sequential spending behavior (Saputra et al., 2019). Recent research highlights the promise of hybrid ML–DL approaches, which combine the predictive power of deep neural networks with the interpretability of tree-based models (Gandhar, Gupta, Pandey, & Raj, 2024). Advanced feature

shengkun.xie@torontomu.ca (S. Xie*)
ORCID(s): 0000-0002-9533-2096 (S. Xie*)

selection frameworks that integrate statistical and evolutionary methods have also been proposed to reduce redundant features, lower computational cost, but retain discriminative power (Siam, Bhowmik, & Uddin, 2025).

DL has emerged as a dominant paradigm in recent years due to its ability to automatically extract high-level representations from raw transaction data (Du, Lv, Guo, & Wang, 2023; Zioviris, Kolomvatsos, & Stamoulis, 2024). Deep neural networks have been shown to scale well to distributed systems and maintain generalization in dynamic environments (PRASAD & Srikanth, 2024). Furthermore, their ability to perform anomaly detection and representation learning makes them highly effective when fraud patterns evolve over time (Mehbodniya et al., 2021). Nevertheless, DL models often require large volumes of labeled data, demand substantial computational resources, and suffer from limited interpretability, posing challenges for deployment in highly regulated financial environments (Mienye, Jere, Obaido, Mienye, & Aruleba, 2024).

Building on these insights, the present study addresses these gaps through a comprehensive investigation of supervised ML approaches for fraudulent detection of financial transactions. Specifically, the study (i) systematically compares seven ML models, including logistic regression, random forests, XGBoost, shallow and deep neural networks, and a novel hybrid ensemble; (ii) evaluates imbalance handling strategies, such as class weighting, focal loss, and threshold optimization; (iii) incorporates transactional, temporal, and identity-linked features to improve contextual predictive power; and (iv) applies SHAP-based interpretability analysis to ensure transparency and regulatory compliance.

The main contributions of this research are threefold: (1) a detailed characterization of fraud patterns across transactional, temporal, and identity dimensions; (2) a hybrid ensemble model integrating Enhanced XGBoost and a Deep Neural Network, achieving state-of-the-art performance on the dataset we consider; and (3) an adaptive framework that combines explainability, threshold tuning, and retraining strategies to maintain long-term resilience. Collectively, this work advances the development of scalable, accurate, and interpretable fraud detection systems suitable for real-world application.

2. Related Work

Research on fraud detection has largely focused on supervised ML approaches, aiming to address the challenges posed by real-world datasets that are highly imbalanced, high-dimensional, complex and often noisy. Existing studies can be broadly classified into four themes: (i) model selection, (ii) imbalance-handling techniques, (iii) feature engineering, and (iv) the integration of identity-based variables to improve predictive power of the model.

Supervised learning models remain fundamental for most modern fraudulent detection systems (Bin Sulaiman, Schetinin, & Sant, 2022; George, Alam, & Hasan, 2025). These models learn parameters or hyper-parameters from labeled historical data and build the model to predict whether new transactions are fraudulent or legitimate. A notable advancement in this area is the distributed deep neural network (DDNN) proposed by Lei et al. (2023), which balances predictive accuracy with user privacy by allowing institutions to train local models and share only model parameters with a central server. This federated-style approach not only preserves data confidentiality but also improves efficiency through distributed computation, yielding superior accuracy, precision, recall, and F1-scores compared to centralized models (Xia & Saha, 2025).

Interpretable models such as decision trees and random forests have also been widely used in fraudulent detection (Lee, Fu, Wang, & Azis, 2025; Sun, 2025; Wajgi, Agarkar, Patil, Rao, & Petkar, 2024). Decision trees are valued for their transparency but are prone to overfitting, whereas random forests mitigate this issue through tree aggregation and hyperparameter tuning (Shah & Sharma, 2023). However, both approaches are sensitive to severe class imbalance. A comprehensive evaluation of 66 algorithm–resampling combinations by Alfaiz and Fati (2022) highlighted that pairing strong classifiers such as CatBoost with effective undersampling methods like AllKNN can achieve state-of-the-art performance in terms of F1-score, recall, and AUC.

Class imbalance remains one of the most critical obstacles in fraudulent detection (Baisholan et al., 2025; Velarde et al., 2023). When fraudulent cases account for less than 1% of transactions, models are significantly biased toward the majority class, leading to poor recall. Numerous resampling and reweighting techniques have been proposed to mitigate this challenge. For example, Alamri and Ykhlef (2024) introduced BCB-SMOTE, a hybrid method combining Tomek links, clustering, and borderline synthetic oversampling, which achieved an F1-score of 85.2% while reducing overlap between classes. Similarly, Ileberi, Sun, and Wang (2021) showed that combining SMOTE with AdaBoost improved detection rates across classifiers, and Lahbiss and Chtouki (2024) reported that SMOTE-ENN with advanced models such as Random Forest and Long Short Term Memory (LSTM) significantly enhanced AUC-ROC. Moreover, Jiao, Guo, Gong, and Chen (2022) proposed DES-ICD, which integrates adaptive oversampling (AnnSMOTE) with dynamic ensemble selection to handle both class imbalance and concept drift. By generating minority samples that reflect new concepts and selecting classifiers based on local neighborhood performance, DES-ICD achieved superior accuracy and recall across multiple real and synthetic datasets.These findings highlight that careful design of imbalance handling techniques is crucial for achieving high recall without sacrificing precision.

Beyond class imbalance, feature engineering plays a pivotal role in fraudulent detection (Alamri & Ykhlef, 2024; Sharma, Sharma, Malik, Sobti, & Suryana, 2025). The inclusion of behavioral and identity-based attributes has been shown to significantly improve model performance. For instance, Shimin, Ke, Xinye, et al. (2020) demonstrated that combining financial transaction data with identity-linked features (e.g., device type, email domain) significantly enhanced

XGBoost's ROC-AUC to 0.942 on the IEEE-CIS dataset. Similarly, Lucas et al. (2019) applied Hidden Markov Models to capture sequential spending behavior, which, when combined with Random Forest, improved precision–recall AUC. Bahnsen, Aouada, Stojanovic, and Ottersten (2016) further extended this line of work by modeling periodic spending patterns with von Mises distributions, yielding a 13% reduction in financial losses. These studies collectively underscore that well-designed feature engineering strategies are essential for building robust and discriminative fraud detection models.

Ensemble and hybrid learning methods represent a growing trend in fraudulent detection research. Carcillo et al. (2021) combined unsupervised anomaly detection with supervised learning, feeding outlier scores as features into classifiers to enable multi-level detection. Dynamic ensemble selection approaches, such as the one proposed by Achakzai and Peng (2023), adaptively choose classifiers based on local competence and consistently outperform static ensembles. Hybrid models have also demonstrated promise; for example, Jahnavi et al. (2024) combined decision trees with logistic regression to achieve 98.1% accuracy for the data they considered, while Chaurasia, Kesharwani, Sharma, Sharma, and Chugh (2024) confirmed that XGBoost paired with data balancing strategies offers superior recall in rare-event detection.

Several comparative reviews have synthesized findings across models and datasets. Patel, Patel, and Patel (2024) reported that while deep neural networks often achieve the highest accuracy (up to 98.9%), simpler models such as logistic regression and Naive Bayes remain competitive due to their interpretability and high efficiency. Similarly, Bhardwaj, Kumar, Verma, Kumar, et al. (2024) found that deep neural networks trained using the Adam optimizer reached 99.4% accuracy on the European credit card dataset and were computationally efficient, making them well-suited for large-scale data implementation.

Taken together, prior work demonstrates that deep learning and advanced ensemble approaches frequently deliver superior predictive performance, yet simpler models retain value for their interpretability, scalability, and ease of application. Despite these advances, few studies have simultaneously addressed all major challenges, including class imbalance, feature selection, interpretability, and integration of identity-based features, within a unified framework. This work seeks to close this gap by systematically comparing multiple supervised ML models, integrating advanced imbalance handling techniques, and leveraging feature engineering to deliver a scalable, interpretable, and robust fraud detection framework.

3. Materials and Methods

This study adopts a systematic and rigorous methodology for evaluating supervised machine learning models in the context of financial fraud detection. The proposed framework is explicitly designed to address two major challenges inherent to large-scale financial transaction data: high dimensionality and severe class imbalance. The methodological design is organized into four sequential phases, data preprocessing, imbalance mitigation, model training with hyperparameter optimization, and post-hoc interpretability analysis. Each phase is implemented to ensure both model robustness and transparency. The following sections provide detailed descriptions of the procedures and techniques applied within each phase.

3.1. Data and Its Description

The data used in this study is the IEEE-CIS fraudulent detection dataset, released through a Kaggle competition in collaboration with Vesta, a global fraud prevention company. This data reflects real-world e-commerce environments with anonymized transaction and identity information, and its very suitable for testing algorithms and computational framework that designed for fraudulent detection of financial transactions. The training set contains over 590,000 records, of which only about 3.5% are labeled as fraudulent, while the test set contains similar features but does not include fraud labels. The dataset can be downloaded from the official Kaggle competition portal (`https://www.kaggle.com/competitions/ieee-fraud-detection/data`).

Two main files were provided for both training and test: `transaction.csv`, which contains transaction-level attributes, and `identity.csv`, which includes device and identity-related variables. These files were merged using the `TransactionID` field to produce a comprehensive view of each transaction. The features can be grouped into several categories, as summarized in Table 1.

Beyond the feature types, it is also important to consider the overall statistical profile of the dataset. Basic descriptive statistics are provided in Table 2. The dataset is highly imbalanced, with fraudulent cases representing only a small fraction of transactions. Transaction amounts vary widely, ranging from a few cents to over $10,000, with a median value of approximately $68. Additionally, many identity-related fields contain substantial missing values, highlighting the challenges inherent in real-world fraud detection problems.

These characteristics emphasize the dual challenges of extreme class imbalance and data high-dimensionality. These insights directly informed the preprocessing strategies and modeling decisions described in the following sections.

3.2. Data Preprocessing

The training dataset was constructed by merging transaction dataset and identity dataset on the TransactionID field, thereby integrating transaction-level payment attributes with device- and identity-related features. This merging step was crucial to capture both behavioral and contextual features that can distinguish fraudulent from legitimate transactions. To address missing data, categorical variables were imputed with the string "missing" so that models could treat absence of information as an additional informative category. Numerical attributes with missing values were retained as NaN, which allows tree-based methods such as XGBoost to handle them.

Table 1
Summary of feature groups in the IEEE-CIS dataset

Feature Group	Description / Examples
Transaction Features	`TransactionAmt`, `TransactionDT`, `ProductCD`
Card Attributes	`card1–card6` (e.g., card type, issuer, category)
Address Codes	`addr1`, `addr2` (geographic location codes)
Engineered Features	`C1–C14`, `D1–D15`, `V1–V339` (anonymized signals)
Email Domains	`P_emaildomain`, `R_emaildomain`
Identity Features	`DeviceType`, `DeviceInfo`, `id12–id38` (browser, OS, network)

Table 2
Descriptive statistics of the IEEE-CIS training dataset

Statistic	Value
Total records	~590,000
Fraud proportion	3.5%
Transaction amount range	\$0.01 – \$10,000+
Median transaction amount	\$68
Number of features (after merge)	434

Categorical features, including ProductCD, card4, and DeviceType, were transformed using label encoding to convert string categories into numerical form while preserving their distinct identities. Although more sophisticated encoders (e.g., target or one-hot encoding) could be applied, label encoding was selected to maintain consistency across a high-dimensional feature space and reduce memory overhead. In addition, a temporal feature, hour_of_day, was derived from the continuous TransactionDT timestamp to capture periodic spending behaviors that may indicate fraud, such as late-night or off-hour activity.

Following data preprocessing, the dataset was partitioned into training and validation subsets using an 80/20 stratified split. Stratification ensured that the proportion of fraudulent to legitimate transactions was maintained in both sets, enabling a fair and representative evaluation. The computational details of this preprocessing and splitting is presented in Algorithm 1. This approach preserved the natural distribution of the data, which is essential for imbalanced classification problems. However, it also retained potential noise from weak or redundant features, meaning that subsequent feature selection and model regularization were critical for improving robustness.

Algorithm 1: Preprocessing and Stratified Split

Input: Transaction file T, Identity file I
Output: $(X_{train}, y_{train}, X_{val}, y_{val})$
Load T and I from CSV files;
Replace hyphens in column names with underscores;
for *each categorical column c in T and I* **do**
    Fill missing values in c with "missing";
Merge T and I on `TransactionID`;
for *each categorical column c in merged dataset* **do**
    Apply Label Encoding to c;
if `TransactionDT` *exists* **then**
    Derive new feature `hour_of_day` $\leftarrow$ (`TransactionDT`//3600) mod 24;
$X \leftarrow$ all features except `isFraud`, `TransactionID`;
$y \leftarrow$ `isFraud`;
Split (X, y) into (X_{train}, y_{train}) and (X_{val}, y_{val}) with 80/20 stratified split;
return $(X_{train}, y_{train}, X_{val}, y_{val})$;

3.3. Handling Class Imbalance

Fraudulent transactions represented only about 3.5% of the dataset, creating a severe class imbalance that posed a significant challenge for model training. If left unaddressed, most classifiers would become biased toward predicting the majority class (legitimate transactions), thereby achieving deceptively high accuracy but failing to identify the rare fraud cases that matter most in practice. To mitigate this imbalance, cost-sensitive learning approaches were adopted. For gradient boosting models such as XGBoost, the parameter value of scale_pos_weight was set to the ratio of majority to minority class instances, thereby instructing the algorithm to assign higher importance to fraud cases during training. For neural networks, a focal loss function was used. Unlike traditional cross-entropy, focal loss dynamically down-weights well-classified examples and focuses learning on harder, misclassified fraud cases. This adaptation is particularly effective when fraudulent behavior exhibits high diversity and overlaps with legitimate transaction patterns. This strategy preserved the natural class distribution of the dataset and avoided the introduction of synthetic artifacts, which are often a drawback of oversampling or SMOTE-based techniques. However, while cost-sensitive methods improve recall, they may not fully resolve imbalance in scenarios where the decision boundary between classes is highly non-linear or overlapping.

3.4. Machine Learning Models

In this section, we provide a brief overview of the machine learning model used in this study to ensure the paper remains self-contained. Five models were selected based on their suitability for handling structured, imbalanced data classification problems and their diversity in algorithmic approach. Let $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ denote a sequence of training set of fi-

nancial transaction data, where $\mathbf{x}_i \in \mathbb{R}^p$ is the p-dimensional feature vector, and $y_i \in \{0, 1\}$ is the class label of fraud or non-fraud. We first consider Logistic Regression (LR), a linear baseline model that estimates the conditional probability of the positive (fraudulent) class as

$$\hat{y}_i = P(y_i = 1 \mid \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i + b),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid function, and $\hat{y}_i \in [0, 1]$ represents the predicted probability of fraud. The unknown parameters $\mathbf{w}$ and b can be estimated using the least squares method or maximum likelihood estimation.

To capture non-linear patterns, we also apply a Random Forest (RF) model, which is an ensemble of T decision trees $\{h_t\}_{t=1}^{T}$. Each tree is trained on a bootstrap sample of the data with random feature subsampling. The final prediction is the average of individual tree predictions of probability,

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}_i),$$

where $h_t(\mathbf{x}_i) \in [0, 1]$ denotes the probability assigned by tree t.

As a more powerful alternative, XGBoost uses gradient boosting to iteratively construct an additive model,

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i),$$

where f_t represents the regression tree added at iteration t. The objective function minimized at each step is

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}) + \Omega(f_t),$$

where $\ell(\cdot)$ is a differentiable loss function, such as logistic loss, and $\Omega(\cdot)$ is a regularization term penalizing tree complexity to prevent overfitting.

We further explore neural network-based approaches. The Artificial Neural Network (ANN) considered here is a shallow feed-forward network with a single hidden layer. For hidden representations computed as $\mathbf{h} = \phi(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1)$, where $\phi(\cdot)$ is an activation function such as ReLU, the predicted probability is given by

$$\hat{y}_i = \sigma(\mathbf{w}_2^\top \mathbf{h} + b_2).$$

Finally, we utilize a Deep Neural Network (DNN) with L hidden layers. The prediction is expressed as

$$\hat{y}_i = \sigma\big(f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}(\mathbf{x}_i)\big),$$

where $f^{(l)}(\cdot)$ represents the non-linear transformation at layer l. To address class imbalance, the DNN is trained using the focal loss function defined as follows

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{n} \alpha(1-\hat{y}_i)^\gamma y_i \log \hat{y}_i + (1-\alpha)\hat{y}_i^\gamma (1-y_i) \log(1-\hat{y}_i),$$

where α balances the class weights and γ down-weights well-classified examples, focusing learning on harder cases.

The inclusion of these models provided both breadth and depth in evaluation. Logistic regression offered transparency and interpretability, serving as a benchmark for assessing gains from more complex models. Random forests and XGBoost captured non-linear interactions through ensemble learning, with XGBoost effectively handling imbalance via class weighting. Neural networks represented higher-capacity learners: the ANN provided a shallow deep learning baseline, while the DNN leveraged deeper architectures and focal loss for improved representation learning in imbalanced fraud detection tasks.

To combine the interpretability and robustness of boosting with the non-linear representational power of deep neural networks, a hybrid ensemble was designed. The model integrates XGBoost feature representations with a DNN, where the outputs of XGBoost are used alongside the original features to provide enriched inputs for the neural network. This design enables the DNN to learn both raw transaction patterns and boosted interaction signals. The overall workflow is summarized in Algorithm 2.

Algorithm 2: Hybrid Ensemble Training (XGBoost + DNN)

Input: X_{train}, y_{train}, X_{val}, y_{val}
Output: Hybrid model H

```
Train XGB on (X_train, y_train) with imbalance
  weight w;
xgb_probs ← XGB.predict(X_val);
Train DNN on (X_train, y_train) with focal loss;
dnn_probs ← DNN.predict(X_val);
ensemble_probs :=
  0.6 · xgb_probs + 0.4 · dnn_probs;
Tune threshold to maximize F1 on y_val;
H := (XGB, DNN, weights, threshold);
return H;
```

In this hybrid approach, XGBoost contributes structured feature learning and interpretability, while the DNN captures non-linear patterns, enabling robust fraud detection in imbalanced and high-dimensional environments. The weights (0.6 for XGBoost, 0.4 for DNN) were selected based on validation performance, giving slightly greater influence to XGBoost due to its stability on tabular data while still leveraging the DNN's capacity to refine decision boundaries. This weighting produced the best balance between recall and precision, ensuring that the ensemble remains both accurate and adaptable for practical deployment.

3.5. Experimental Design

The methodological framework discussed in previous Sections laid the foundation for a structured experimental design. Having defined the preprocessing pipeline, imbalance handling strategies, and predictive models, the next step was to implement these methods in a controlled series of experiments. These experiments were carefully structured to test models of varying complexity under realistic fraud detection conditions.

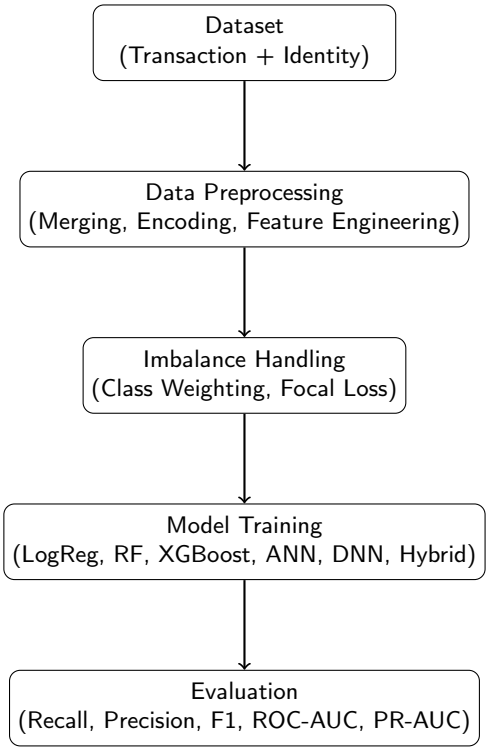


Figure 1: Staged workflow of the experimental design, from dataset preparation to evaluation

3.5.1. Objectives and Rationale

The primary objective of the experiments was to evaluate the effectiveness of various supervised machine learning models for detecting fraudulent transactions using the IEEE-CIS dataset. Three considerations guided the experimental design. First, simple and interpretable models were assessed as baseline approaches. Second, ensemble methods, including Random Forest and XGBoost, were examined for their ability to improve discrimination under severe class imbalance. Third, deep neural networks and hybrid ensembles were investigated to capture non-linear patterns potentially missed by tree-based methods.

This approach reflects the operational priorities of fraud detection, which require a balance between predictive accuracy, interpretability, scalability, and computational efficiency. The experiments followed a staged workflow: dataset preparation and preprocessing, class imbalance mitigation, evaluation of increasingly complex model families, and performance assessment using metrics appropriate for imbalanced classification. The overall workflow and the range of models are illustrated in Figure 1.

3.5.2. Models Under Evaluation

Six models, representing distinct methodological families, were selected to evaluate performance across the comparative framework. Traditional baselines and ensemble learners included Logistic Regression as a transparent, low-cost benchmark; Random Forest, which captures non-linear interactions and handles missing data effectively; and XGBoost, a gradient boosting algorithm designed for high performance on imbalanced data through iterative boosting, regularization, and class weighting. Neural network models comprised a shallow ANN to assess whether modest architectures could detect fraud patterns beyond tree-based methods, and a deeper DNN with additional hidden layers, dropout regularization, and focal loss to enhance representation and handle class imbalance. Finally, a hybrid ensemble combined XGBoost and DNN. Here, XGBoost generated probability scores and feature importance rankings, which were appended to the original features and used as inputs to the DNN. This integration leveraged both structured feature learning and deep representations, aiming to improve predictive performance in real-world fraud detection.

3.5.3. Evaluation Framework

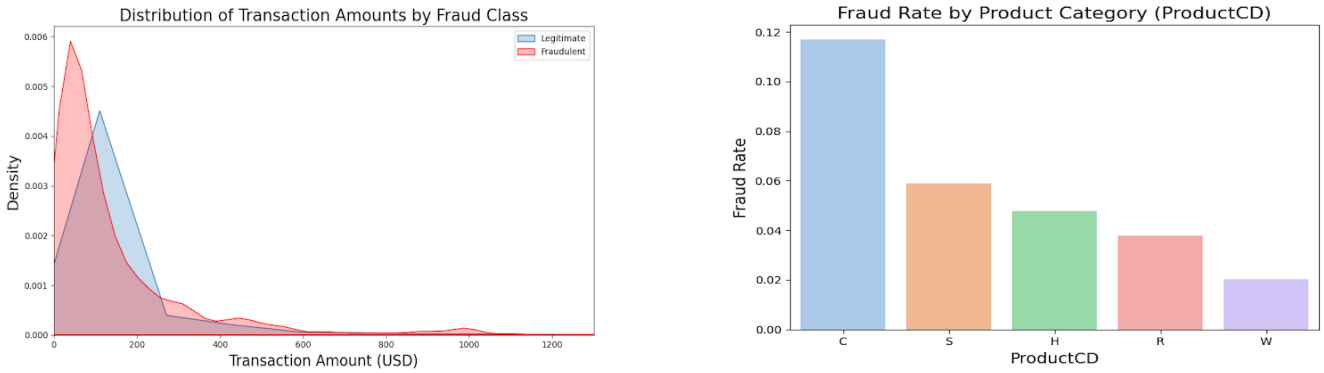
Model evaluation in fraud detection requires careful consideration of extreme class imbalance. Overall accuracy was excluded due to its tendency to overstate performance in majority-dominated datasets. Instead, a combination of complementary metrics was employed to assess both detection effectiveness and operational impact. Recall (sensitivity) was prioritized to minimize costly false negatives, while precision ensured that gains in recall did not lead to excessive false alarms. The F1-score provided a balanced measure of this trade-off. Discriminative ability was captured via ROC-AUC, with additional emphasis on PR-AUC, which better reflects performance on the minority (fraudulent) class. Finally, a cost-sensitive adjustment penalized false negatives more heavily than false positives, aligning evaluation with the financial consequences of undetected fraud. This multi-metric framework ensured models were assessed for both predictive accuracy and practical utility.

4. Results

This section integrates both exploratory data analysis and modeling, providing a unified presentation of findings. Exploratory data analysis is incorporated directly into the results to illustrate how transaction characteristics and feature distributions inform subsequent model performance. The subsections are structured to highlight specific aspects of the data and connect them to fraud detection outcomes.

4.1. Fraud Distribution and Class Imbalance

The first step in characterizing the dataset was to assess the distribution of fraudulent versus legitimate transactions. The dataset is heavily imbalanced: fraudulent cases represent less than 0.2% of all transactions. This extreme imbalance underscores why conventional accuracy is an unreliable metric in this domain. A trivial classifier that always predicts the majority (non-fraud) class would achieve nearly 99.8% accuracy, yet it would fail to detect any fraudulent behavior. The imbalance also highlights the operational challenge of fraud detection. The imbalance-aware framing of results ensures that subsequent model comparisons align with practical fraud detection objectives.



(a) Transaction amount distribution by fraud class.

(b) Fraud rate by product category.

Figure 2: Fraud patterns across transaction amounts and product categories.

4.2. Transaction and Product Patterns

Fraudulent activity often manifests through systematic behaviors rather than random noise, making transaction characteristics an important source of discriminative features. In particular, monetary value and product type reveal clear differences between fraudulent and legitimate transactions, as illustrated in Figure 2. Figure 2a presents the distribution of transaction amounts by class. Fraudulent activity is disproportionately concentrated in very low value transactions, often below 200 USD, which suggests that fraudsters frequently conduct small "test" purchases to confirm the validity of stolen credentials before escalating to larger transactions. At the other extreme, there is also a visible concentration of fraud in very high value purchases, reflecting opportunistic attempts to maximize financial gain once an account has been compromised. Legitimate transactions, by contrast, are more evenly spread across the range but cluster most heavily in the mid value segment, particularly below 500 USD, after which their frequency declines sharply. This divergence between fraudulent and legitimate spending behaviors highlights the deliberate strategies used by fraudsters to balance concealment and profitability. Figure 2b shows fraud prevalence across product categories. Fraud rates are highest in category C, followed by category S, whereas categories H, R, and especially W exhibit much lower levels of fraudulent activity. These discrepancies align with differences in product risk profiles: categories dominated by digital or card not present transactions are more vulnerable to exploitation, while those tied to physical goods or requiring stronger verification demonstrate greater resilience. Incorporating such product level distinctions into predictive models provides highly discriminative signals that enhance the effectiveness of both ensemble and neural network based approaches to fraud detection.

4.3. Temporal and Geolocation Patterns of Fraud

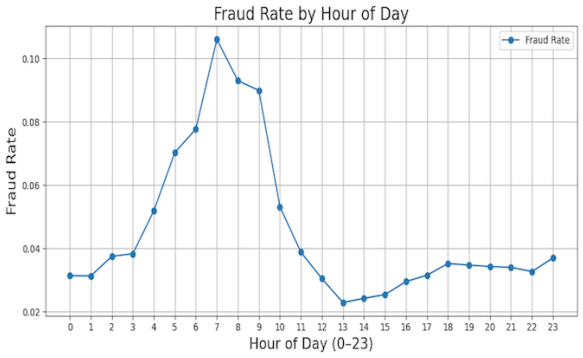
Fraudulent transactions also exhibit systematic temporal and geographical behaviors rather than occurring uniformly across time and space. These patterns, summarized in Figure 3, reveal how fraud risk is influenced by daily cycles, weekly rhythms, and regional contexts. Figure 3a shows fraud distribution across hours of the day. Fraudulent activity peaks during late-night and early-morning hours, a pattern consistent with reduced customer vigilance and lower institutional monitoring during off-peak periods. Figure 3b illustrates the weekly cycle: fraud rates are elevated on weekends, reflecting opportunistic exploitation of reduced transaction monitoring and weaker operational staffing. Figure 3c and 3d highlight geographical differences. Fraud rates vary significantly across regions and countries, with certain locations showing disproportionately high prevalence. These discrepancies may reflect both differences in fraudster targeting strategies and variability in regional payment infrastructures. Such spatial heterogeneity reinforces the value of incorporating geolocation features into fraud detection models, as they provide powerful discriminative signals when combined with transaction and identity attributes.

4.4. Identity-Based Feature Insights

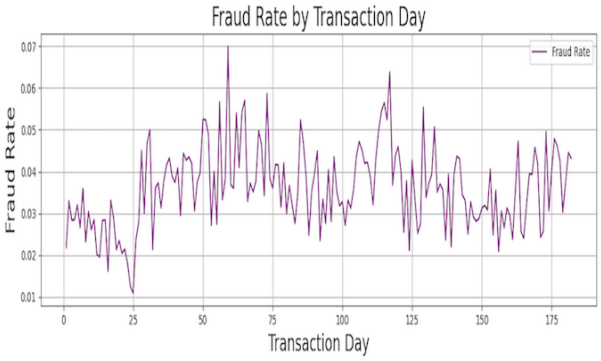
Identity-related features provide some of the strongest discriminative signals in fraud detection, and Figure 4 highlights three representative patterns. Figure 4a shows that free email domains are disproportionately associated with fraudulent activity, whereas institutional or corporate domains exhibit much lower fraud rates, reflecting fraudsters' preference for anonymous or disposable services. Figure 4b compares fraud rates across device types, with mobile and tablet transactions showing higher fraud levels than desktop, consistent with weaker authentication mechanisms and less reliable device fingerprinting. Figure 4c illustrates variation by device information, where certain identifiers appear disproportionately in fraudulent transactions, suggesting the use of emulated or spoofed devices or the systematic reuse of compromised profiles. Together, these findings underscore why identity-linked attributes are heavily weighted by both tree-based ensembles and deep learning models in fraudulent detection.

4.5. Results of Model Performance

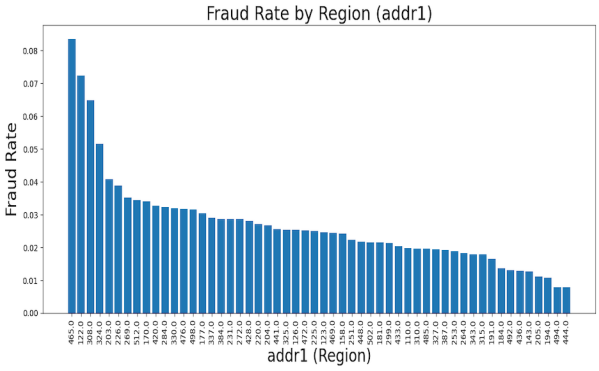
This section presents a comparative evaluation of all seven machine learning models developed for transaction fraud detection: Logistic Regression (LR), Random Forest (RF), XG-



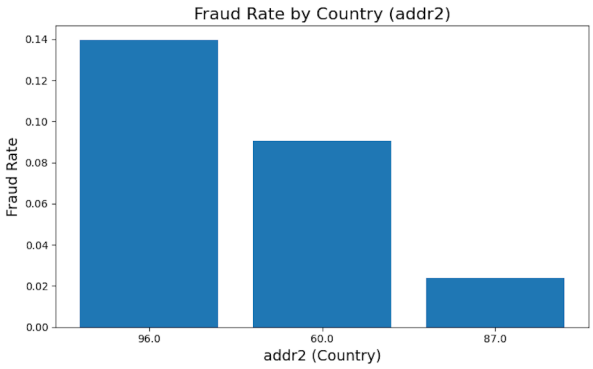
(a) Fraud distribution by hour of day.



(b) Fraud distribution by day of week.



(c) Fraud distribution by region.



(d) Fraud distribution by country.

Figure 3: Temporal and spatial fraud patterns: (a) hour of day; (b) day of week; (c) region; (d) country.

Table 3
Comparative performance of models on fraud detection task.

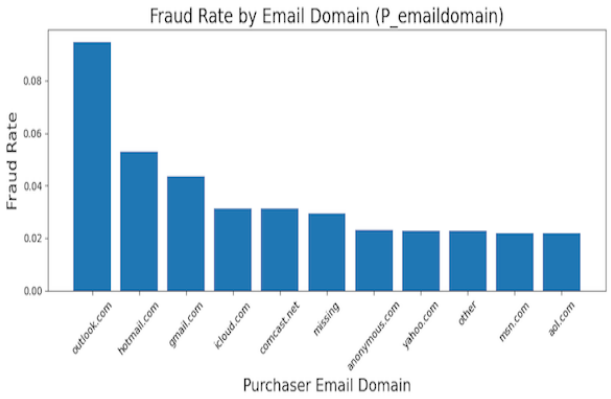
Model	Accuracy	Recall	Precision	F1-Score	ROC-AUC	PR-AUC
Logistic Regression (LR)	77%	0.69	0.10	0.18	0.7959	0.1772
Random Forest (RF)	94%	0.68	0.33	0.45	0.9122	0.5856
XGBoost (XGB)	89%	0.81	0.22	0.35	0.9276	0.6187
Enhanced XGBoost (XGB-Adv)	95%	0.83	0.42	0.55	0.9628	0.7676
Artificial Neural Network (ANN)	98%	0.43	0.91	0.59	0.9160	0.6403
Deep Neural Network (DNN)	98%	0.51	0.82	0.63	0.9182	0.6582
XGB + DNN Ensemble	98%	0.69	0.78	0.74	0.9638	0.7897

Boost (XGB), Enhanced XGBoost (XGB-Adv), Artificial Neural Network (ANN), Deep Neural Network (DNN), and the XGBoost + DNN Ensemble. Table 3 summarizes the validation performance of all models. Metrics include ROC-AUC and PR-AUC for ranking capability, as well as accuracy, precision, recall, and F1-score for fraud detection effectiveness.

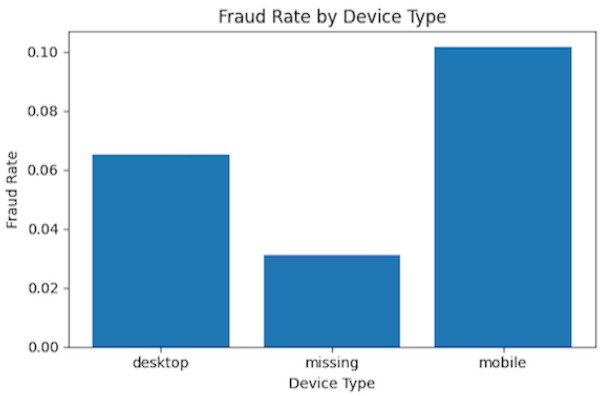
The results show a clear performance gap between baseline models and advanced approaches. Logistic Regression suffers from low precision (0.10) and F1-score (0.18), making it unsuitable for fraud detection. Random Forest improves on these metrics, reaching an F1-score of 0.45, but still falls behind gradient boosting methods in capturing complex relationships in the data. Standard XGBoost raises recall to 0.81, boosting fraud capture rates, but precision (0.22) remains low, leading to a higher false-positive rate.

Enhanced XGBoost demonstrates the most significant improvement among single models, delivering the highest recall (0.83) and substantially better precision (0.42) than standard XGBoost. This balance results in a higher F1-score (0.55) and strong PR-AUC (0.7676), reflecting better performance in the imbalanced fraud detection setting. The improvement can be attributed to careful hyperparameter tuning and regularization, which reduce overfitting while improving fraud detection sensitivity. For use cases prioritizing maximum fraud capture with acceptable false positives, Enhanced XGBoost is a strong candidate.

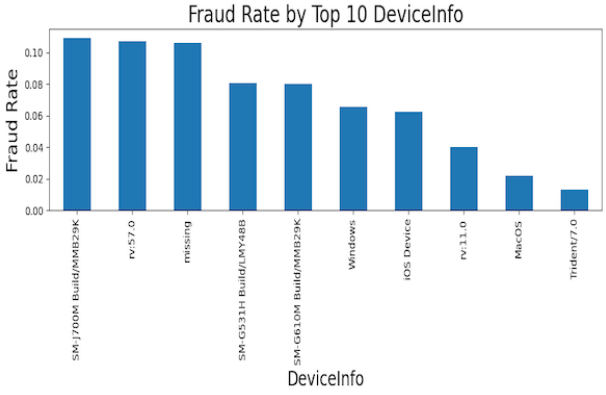
The XGB + DNN Ensemble outperforms all other models in overall balance, achieving the highest F1-score (0.74) and PR-AUC (0.7897), alongside a high recall (0.69) and precision (0.78). By combining Enhanced XGBoost's struc-



(a) Fraud distribution by email domain.



(b) Fraud distribution by device type.



(c) Fraud distribution by device information.

Figure 4: Identity-based fraud patterns: (a) email domain; (b) device type; (c) device information.

Table 5: ROC-AUC comparison between the proposed models and the OLightGBM.

Model	ROC-AUC
Enhanced XGBoost (XGB-Adv)	0.9628
XGB + DNN Ensemble	0.9638
OLightGBM (Taha & Malebary)	0.9288

tured feature learning with the DNN's deep representation capability, the ensemble reduces weaknesses present in each standalone model. This makes it well-suited for real-world usage, where both detecting fraud and minimizing false positives are critical for efficiency and customer trust.

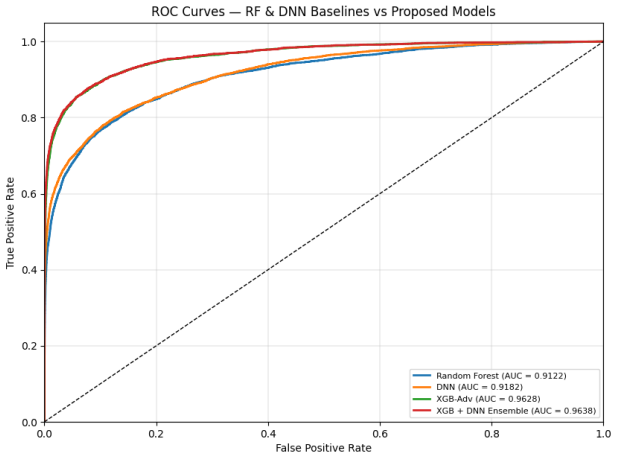
4.6. External Benchmarking

To further assess the competitiveness of the proposed models, the Enhanced XGBoost and the XGBoost + DNN Ensemble was compared against the OLightGBM model, which serves as a strong published benchmark for fraud detection in similar research areas. The comparison focuses on the ROC-AUC metric, which provides an overall measure of a model's discriminative ability across all classification thresholds.

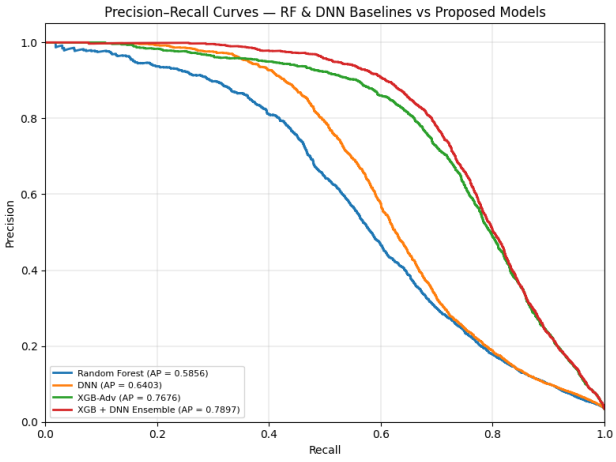
Both proposed models outperform the OLightGBM benchmark in terms of ROC-AUC, demonstrating their competitive advantage in fraud detection tasks. The XGB + DNN Ensemble achieved the highest ROC-AUC score of 0.9638, surpassing the Enhanced XGBoost model at 0.9628, and both showing a notable improvement over the OLightGBM score of 0.9288. These results highlight the robustness of the proposed approaches and their potential to deliver more reliable fraud detection in real-world applications.

4.7. ROC and Precision–Recall Curve Analysis

This section compares four models to evaluate how the proposed approaches perform against strong baselines. Random Forest is included as a representative of traditional machine learning methods, while the Deep Neural Network represents deep learning approaches. These are compared with the Enhanced XGBoost (XGB-Adv) and the XGB + DNN Ensemble, which combine boosting and neural networks. This comparison highlights the improvements of the proposed methods over both classical and deep learning models. Figure 6 presents the ROC and Precision–Recall (PR) curves for the four models. Both plots confirm the superior performance of the proposed methods over the baselines. In Figure 5a, the XGB + DNN Ensemble achieved the highest ROC-AUC of 0.9638, followed closely by XGB-Adv at 0.9628. Among the baselines, the DNN (0.9182) slightly outperformed the



(a) ROC curves for RF, DNN, XGB-Adv, and XGB + DNN Ensemble.



(b) PR curves for RF, DNN, XGB-Adv, and XGB + DNN Ensemble.

Figure 5: ROC and Precision–Recall curve analysis for baseline and proposed models: (a) ROC curves; (b) PR curves.

RF (0.9122). The steep initial slope and proximity of the proposed models' curves to the top-left corner illustrate their strong discriminative ability across thresholds. Figure 5b shows the PR curves, which are more informative for imbalanced datasets. The XGB + DNN Ensemble again outperformed all models with an Average Precision (AP) of 0.7897, followed by XGB-Adv at 0.7676. Among the baselines, the DNN achieved 0.6403, outperforming the RF at 0.5856. The proposed models maintain higher precision at varying recall levels, demonstrating superior handling of the minority fraud class.

The results demonstrate that while traditional models such as Logistic Regression and Random Forest provide useful baselines, they fall short under severe class imbalance. Enhanced XGBoost achieved substantial gains, and the hybrid XGB + DNN Ensemble consistently outperformed all other approaches, delivering the highest recall, F1-score, ROC-AUC, and PR-AUC. Exploratory analyses further revealed clear fraud patterns across transaction amounts, product categories, temporal cycles, and identity-related features, confirming their importance as discriminative signals. Taken together, these findings underscore both the necessity of advanced ensemble methods and the value of incorporating domain-specific patterns, providing a comprehensive foundation for fraud detection pipelines.

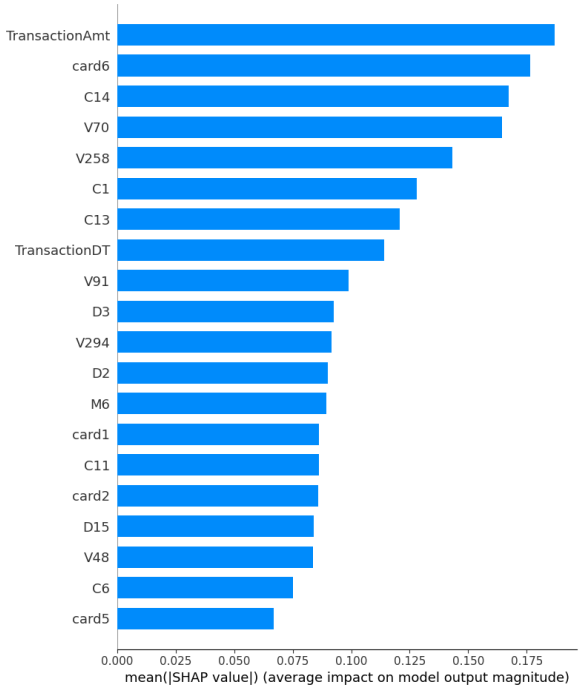


Figure 6: Global Feature Importance - Bar Plot

4.8. Results on Feature Interpretability

This section extends the analysis beyond performance metrics to interpret the results and elucidate their practical significance. While the preceding section identified the models that performed best under class imbalance, the present discussion focuses on understanding the underlying mechanisms driving their success, the insights revealed by key features, and the implications for real-world fraud detection. Model interpretability remains essential in financial applications, where transparency and accountability are critical for institutional deployment. Accordingly, SHAP (SHapley Additive exPlanations) values were computed for the Enhanced XGBoost model—the best-performing single learner in this study, to quantify both global and local feature contributions. This analysis provides a deeper understanding of the predictive signals and their relative importance in detecting fraudulent transactions.

4.8.1. Global Feature Importance

The SHAP global importance plot (see Figure 6) ranks features by their overall contribution across validation data. Transaction amount (`TransactionAmt`) emerged as the most

influential predictor: unusually small or large amounts consistently raise the probability of fraud. Identity-related features such as `card6` (card type) and anonymized variables (`C14`, `C1`, `C13`) were also highly predictive, likely capturing structural patterns in customer identity verification. Engineered behavioral features (e.g., `V70`, `V258`, `V91`, `V294`) and temporal variables such as `TransactionDT` and `D*` deltas further indicate that fraudulent behavior is closely tied to unusual timing anomalies and device usage. These findings highlight that effective fraud detection requires a multidimensional view: combining transaction size, identity markers, and behavioral timing signals. The model's reliance on diverse feature groups suggests that narrowly defined heuristics would fail to capture the full spectrum of fraud activity.

4.8.2. Local Explanation

The SHAP beeswarm plot shown in Figure 7 provides transaction-level insights. Red points on the right side of the distribution indicate feature values that push predictions toward fraud, while blue points on the left reduce the fraud score. For example, high transaction amounts almost always increase predicted fraud probability, whereas small amounts suppress it. Features like `card6` and `C14` show conditional effects: certain values elevate risk, while others are benign. These local explanations are particularly valuable for analysts. They provide a case-by-case rationale for why a transaction was flagged, enabling more efficient investigation and ensuring that fraud detection is not a "black box." This interpretability fosters regulatory compliance and builds trust in applying advanced machine learning systems in financial problems.

4.8.3. Comparison with Random Forest

For comparison, the Random Forest feature importance plot (See Figure **??**) shows a broader distribution of influential variables rather than dominance by a small subset. Features such as `C14` and `C13` emerge as the strongest contributors, while transaction-based attributes like `TransactionAmt`, `V258`, and `V265`, along with several V-series variables, also rank highly. This balance indicates that Random Forest relies on both identity signals and behavioral transaction features, capturing multiple dimensions of fraud risk. Importances are more evenly spread across categories, suggesting that no single variable alone drives the predictions but rather a combination of complementary signals.

While this confirms the relevance of key features, Random Forest importances provide only relative weights and lack the ability to explain whether a feature increases or decreases the likelihood of fraud. They also do not account for interactions between features or variation across individual cases. These limitations highlight why more interpretable methods such as SHAP are better suited for high-stakes fraud detection, where fine-grained reasoning and case-level explanations are essential for operational use.

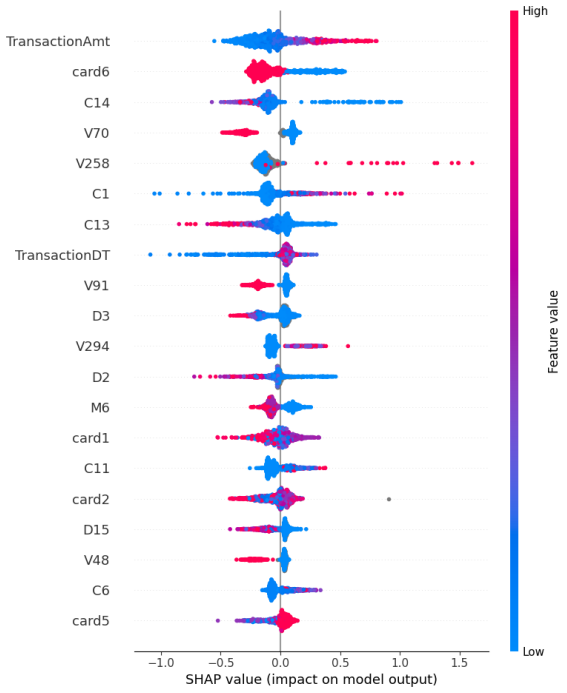


Figure 7: Local Explanation - Beeswarm Plot

5. Discussions

The results carry direct implications for real-world fraud detection systems, where predictive performance must be balanced with operational considerations. Model selection cannot rely solely on ROC-AUC or PR-AUC values; institutions must also consider interpretability, latency, and the trade-off between false positives and false negatives.

While this study demonstrates the effectiveness of advanced ensemble models for fraud detection, several limitations should be acknowledged. First, the dataset, although large and representative of real-world payment transactions, is anonymized and lacks certain contextual features (e.g., merchant category codes, customer demographics, and real-time session information). The absence of these variables restricts the interpretability of fraud patterns and may limit the generalizability of the findings across different financial institutions and geographies.

Second, the evaluation was conducted in an offline validation setting. Although metrics such as ROC-AUC and PR-AUC are informative, they do not capture all operational trade-offs. Real-world deployment involves latency constraints, streaming data pipelines, and integration with fraud investigation teams. These aspects were not simulated in this study, and future research should extend validation to production-like environments, including stress testing under high transaction volumes and adversarial attack scenarios.

Third, while SHAP provided valuable explainability for tree-based models, interpretability for deep neural networks remains less developed. Methods such as LIME or integrated gradients offer partial insights, but their stability and regulatory acceptance are still evolving. This creates chal-

lenges for deploying complex models in strictly regulated financialdomains where transparency is a non-negotiable requirement.

6. Conclusion

This study addressed the persistent challenge of fraud detection in highly imbalanced financial transaction data by systematically comparing traditional, ensemble-based, and deep learning approaches. Through rigorous experimentation, it was demonstrated that classical baselines such as Logistic Regression and Random Forest, while interpretable, fail to provide the precision and recall balance required for real-world applications. Enhanced XGBoost improved performance by leveraging gradient boosting with advanced handling of imbalance, yet it was the proposed hybrid XGB + DNN Ensemble that delivered the most effective and robust results across all evaluation metrics. Beyond raw performance scores, the study emphasized the importance of interpretability and operational readiness. SHAP-based feature attribution highlighted the central role of transaction amounts, identity-linked attributes, and temporal patterns in detecting fraudulent behavior, underscoring the need for multidimensional representations of customer activity.

The hybrid XGB + DNN Ensemble, embedded within a structured and adaptive framework, represents a powerful and practical solution for modern fraud detection. By aligning methodological innovation with operational realities, this study demonstrates that effective fraud detection requires not only strong algorithms but also transparent, adaptive, and institutionally aligned systems capable of withstanding the evolving tactics of financial fraud. Together, these findings advance both academic understanding and practical implementation of fraud detection systems. Nevertheless, limitations remain. The anonymized dataset constrained interpretability of certain fraud patterns, and evaluation was restricted to offline validation. Addressing these gaps through richer feature sets, real-time testing, and advanced learning paradigms such as graph-based or federated learning offers promising directions for future research.

References

Achakzai, M. A. K., & Peng, J. (2023). Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis*, *89*, 102827.

Alamri, M., & Ykhlef, M. (2024). Hybrid feature engineering based on customer spending behavior for credit card anomaly and fraud detection. *Electronics*, *13*(20), 3978.

Alfaiz, N. S., & Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. *Electronics*, *11*(4), 662.

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, *51*, 134–142.

Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). A systematic review of machine learning in credit card fraud detection under original class imbalance. *Computers*, *14*(10), 437.

Bhardwaj, K., Kumar, M., Verma, R., Kumar, D., et al. (2024). Machine learning and deep learning for credit card fraud detection: A comparative analysis. In *2024 international conference on artificial intelligence and emerging technology (global ai summit)* (pp. 131–136).

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.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, *17*(3), 235–255.

Breskuvienė, D., & Dzemyda, G. (2024). Enhancing credit card fraud detection: highly imbalanced data case. *Journal of Big Data*, *11*(1), 182.

Broby, D. (2021). Financial technology and the future of banking. *Financial Innovation*, *7*(1), 47.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, *557*, 317–331.

Chaurasia, S., Kesharwani, S., Sharma, S., Sharma, S., & Chugh, B. (2024). Analysis of ensemble machine learning models for fraud detection. In *2024 international conference on intelligent systems for cybersecurity (iscs)* (pp. 1–6).

Ding, N., Ruan, X., Wang, H., & Liu, Y. (2025). Automobile insurance fraud detection based on pso-xgboost model and interpretable machine learning method. *Insurance: Mathematics and Economics*, *120*, 51–60.

Du, H., Lv, L., Guo, A., & Wang, H. (2023). Autoencoder and lightgbm for credit card fraud detection problems. *Symmetry*, *15*(4), 870.

Gandhar, A., Gupta, K., Pandey, A. K., & Raj, D. (2024). Fraud detection using machine learning and deep learning. *SN Computer Science*, *5*(5), 453.

George, M. Z. H., Alam, M. K., & Hasan, M. T. (2025). Machine learning for fraud detection in digital banking: a systematic literature review review. *arXiv preprint arXiv:2510.05167*.

Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost. *IEEE access*, *9*, 165286–165294.

Jahnavi, D., Mona, A., Pulata, S., Sami, S., Vakamullu, B., et al. (2024). Robust hybrid machine learning model for financial fraud detection in credit card transactions. In *2024 2nd international conference on intelligent data communication technologies and internet of things (idciot)* (pp. 680–686).

Jiao, B., Guo, Y., Gong, D., & Chen, Q. (2022). Dynamic ensemble selection for imbalanced data streams with concept drift. *IEEE transactions on neural networks and learning systems*, *35*(1), 1278–1291.

Lahbiss, M. M., & Chtouki, Y. (2024). Credit card fraud detection in imbalanced datasets: A comparative analysis of machine learning techniques. In *2024 international conference on computer and applications (icca)* (pp. 1–6).

Lee, C.-W., Fu, M.-W., Wang, C.-C., & Azis, M. I. (2025). Evaluating machine learning algorithms for financial fraud detection: insights from indonesia. *Mathematics*, *13*(4), 600.

Lei, Y.-T., Ma, C.-Q., Ren, Y.-S., Chen, X.-Q., Narayan, S., & Huynh, A. N. Q. (2023). A distributed deep neural network model for credit card fraud detection. *Finance Research Letters*, *58*, 104547.

Lucas, Y., Portier, P.-E., Laporte, L., Calabretto, S., Caelen, O., He-Guelton, L., & Granitzer, M. (2019). Multiple perspectives hmm-based feature engineering for credit card fraud detection. In *Proceedings of the 34th acm/sigapp symposium on applied computing* (pp. 1359–1361).

Ludera, D. T. (2021). Credit card fraud detection by combining synthetic minority oversampling and edited nearest neighbours. In *Future of information and communication conference* (pp. 735–743).

Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K. P., Shabaz, M., & Madhavan, M. V. (2021). [retracted] financial fraud detection in healthcare using machine learning and deep learning techniques. *Security and Communication Networks*, *2021*(1), 9293877.

Mienye, E., Jere, N., Obaido, G., Mienye, I. D., & Aruleba, K. (2024). Deep learning in finance: A survey of applications and techniques. *AI*, *5*(4), 2066–2091.

Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., &

Idroes, R. (2023). Credit card fraud detection for contemporary financial management using xgboost-driven machine learning and data augmentation techniques. *Indatu Journal of Management and Accounting*, *1*(1), 29–35.

Patel, A., Patel, M., & Patel, P. (2024). Exploring supervised machine learning techniques for detecting credit card fraud: An investigative review. In *Itm web of conferences* (Vol. 65, p. 03006).

PRASAD, M., & Srikanth, T. (2024). Multi-entity real-time fraud detection system using machine learning: Improving fraud detection efficiency using frost-enhanced oversampling.

Riskiyadi, M. (2024). Detecting future financial statement fraud using a machine learning model in indonesia: a comparative study. *Asian Review of Accounting*, *32*(3), 394–422.

Roseline, J. F., Naidu, G., Pandi, V. S., alias Rajasree, S. A., & Mageswari, N. (2022). Autonomous credit card fraud detection using machine learning approach. *Computers and Electrical Engineering*, *102*, 108132.

Saputra, A., et al. (2019). Fraud detection using machine learning in e-commerce. *International Journal of Advanced Computer Science and Applications*, *10*(9).

Shah, D., & Sharma, L. K. (2023). Credit card fraud detection using decision tree and random forest. In *Itm web of conferences* (Vol. 53, p. 02012).

Sharma, A., Sharma, S., Malik, A., Sobti, R., & Suryana, A. (2025). Dynamic feature engineering for adaptive fraud detection. *Engineering Proceedings*, *107*(1), 68.

Shimin, L., Ke, X., Xinye, S., et al. (2020). An xgboost based system for financial fraud detection. In *E3s web of conferences* (Vol. 214, p. 02042).

Siam, A. M., Bhowmik, P., & Uddin, M. P. (2025). Hybrid feature selection framework for enhanced credit card fraud detection using machine learning models. *PLoS One*, *20*(7), e0326975.

Sun, J. (2025). Decision tree-based credit card fraud detection system: Design and optimization. *Economics & Management Information*, 1–5.

Talukder, M. A., Khalid, M., & Uddin, M. A. (2024). An integrated multi-stage ensemble machine learning model for fraudulent transaction detection. *Journal of Big Data*, *11*(1), 168.

Theodorakopoulos, L., Theodoropoulou, A., Tsimakis, A., & Halkiopoulos, C. (2025). Big data-driven distributed machine learning for scalable credit card fraud detection using pyspark, xgboost, and catboost. *Electronics*, *14*(9), 1754.

Velarde, G., Sudhir, A., Deshmane, S., Deshmunkh, A., Sharma, K., & Joshi, V. (2023). Evaluating xgboost for balanced and imbalanced data: application to fraud detection. *arXiv preprint arXiv:2303.15218*.

Wajgi, R., Agarkar, H., Patil, R., Rao, H., & Petkar, N. (2024). Enhancing credit card transaction fraud detection with random forest and robust scaling. In *Aip conference proceedings* (Vol. 3188, p. 040013).

Xia, Z., & Saha, S. C. (2025). Fingraphfl: Financial graph-based federated learning for enhanced credit card fraud detection. *Mathematics*, *13*(9), 1396.

Zhang, Y.-L., Zhou, J., Zheng, W., Feng, J., Li, L., Liu, Z., … others (2019). Distributed deep forest and its application to automatic detection of cash-out fraud. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(5), 1–19.

Zioviris, G., Kolomvatsos, K., & Stamoulis, G. (2024). An intelligent sequential fraud detection model based on deep learning. *The Journal of Supercomputing*, *80*(10), 14824–14847.

ORCID Information

Shengkun Xie

https://orcid.org/0000-0002-9533-2096

Highlights

- *Hybrid ensemble of Enhanced XGBoost and DNN achieves state-of-the-art results.*
- *Imbalance-aware Class-weighting and SHAP improve recall and ensure explainability.*
- *Systematic evaluation framework validates robustness under severe class imbalance.*

Declaration of Interest Statement:

The authors declare that there is no conflict of interest.