computers

MDPI

Article

FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets

Nazerke Baisholan [1,2,*], J. Eric Dietz [3], Sergiy Gnatyuk [4], Mussa Turdalyuly [2,5,*], Eric T. Matson [3] and Karlygash Baisholanova [1]

[1] Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; baisholanova.k@gmail.com
[2] Software Engineering Department, International Engineering and Technological University, Almaty 050060, Kazakhstan
[3] Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA; jedietz@purdue.edu (J.E.D.); ematson@purdue.edu (E.T.M.)
[4] Faculty of Computer Science and Technology, State University "Kyiv Aviation Institute", 03058 Kyiv, Ukraine; s.gnatyuk@kai.edu.ua
[5] School of Digital Technologies, Narxoz University, Almaty 050035, Kazakhstan
* Correspondence: nbaishol@purdue.edu (N.B.); m.turdalyuly@gmail.com (M.T.)

Abstract: Credit card fraud detection is a critical research area due to the significant financial losses and security risks associated with fraudulent activities. This study presents FraudX AI, an ensemble-based framework addressing the challenges in fraud detection, including imbalanced datasets, interpretability, and scalability. FraudX AI combines random forest and XGBoost as baseline models, integrating their results by averaging probabilities and optimizing thresholds to improve detection performance. The framework was evaluated on the European credit card dataset, maintaining its natural imbalance to reflect real-world conditions. FraudX AI achieved a recall value of 95% and an AUC-PR of 97%, effectively detecting rare fraudulent transactions and minimizing false positives. SHAP (Shapley additive explanations) was applied to interpret model predictions, providing insights into the importance of features in driving decisions. This interpretability enhances usability by offering helpful information to domain experts. Comparative evaluations of eight baseline models, including logistic regression and gradient boosting, as well as existing studies, showed that FraudX AI consistently outperformed these approaches on key metrics. By addressing technical and practical challenges, FraudX AI advances fraud detection systems with its robust performance on imbalanced datasets and its focus on interpretability, offering a scalable and trusted solution for real-world financial applications.

Keywords: credit card fraud detection; machine learning; ensemble models; imbalanced datasets; SHAP; anomaly detection; AUC-PR

Academic Editor: Paolo Bellavista

Received: 10 February 2025
Revised: 21 March 2025
Accepted: 24 March 2025
Published: 25 March 2025

Citation: Baisholan, N.; Dietz, J.E.; Gnatyuk, S.; Turdalyuly, M.; Matson, E.T.; Baisholanova, K. FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets. *Computers* **2025**, *14*, 120. https://doi.org/10.3390/computers14040120

1. Introduction

Credit card fraud remains a persistent challenge for financial institutions, requiring advanced machine learning (ML) techniques to detect fraudulent transactions. Credit card fraud involves the unauthorized use of payment cards to conduct illegal monetary transactions, often aimed at obtaining goods, services, or transferring money without the cardholder's permission [1]. As online transactions continue to rise, fraudsters are constantly developing more sophisticated attack strategies, making fraud detection a critical area of research. According to Juniper Research, global losses due to online payment fraud are expected to reach USD 91 billion in 2028, demonstrating the urgent need for

robust and scalable fraud detection systems [2]. Traditional rule-based fraud detection systems struggle to adapt to evolving fraud patterns and often generate high false positive rates, causing financial and operational inefficiencies. In response to these challenges, ML techniques, with their ability to analyze large datasets and identify subtle patterns, are well suited for addressing the limitations of traditional fraud detection systems [3]. ML models can detect anomalies in transaction data that may signify fraudulent activity and offer the ability to continuously learn and adapt to new fraudulent behaviors [4]. This capability makes them an invaluable tool for real-time fraud detection and prevention, as they can dynamically adapt to emerging threats and evolving fraud patterns. Popular approaches include artificial neural networks (ANNs), logistic regression (LR), decision trees (DT), and support vector machines (SVMs), which are well suited for identifying patterns in large datasets.

However, despite these advances, there are still ongoing challenges. One of the primary challenges in fraud detection is the extreme class imbalance in datasets, where fraudulent transactions constitute only a tiny fraction of all transactions. This imbalance can degrade model performance, resulting in a trade-off between overall accuracy and the ability to detect fraudulent cases. Various strategies, such as oversampling, undersampling, and synthetic data generation, have been employed to mitigate this issue [5,6]. However, these methods often introduce biases or fail to generalize effectively in real-world applications.

Another critical issue in fraud detection is the lack of interpretability in predictive models. While deep learning techniques have demonstrated high accuracy in various domains, they are often criticized as "black-box" models, making it difficult for financial institutions and regulators to trust their decision-making process. Explainable AI approaches, such as Shapley additive explanations (SHAP), have gained attention for providing transparency in model predictions by explaining which features contribute most to fraud classification [7,8].

Several recent studies have explored ensemble learning methods to enhance fraud detection performance [9–11]. However, most of these studies focus on balanced datasets, which may not reflect real-world fraud detection scenarios where class imbalance is inherent.

To address these challenges, our work proposes FraudX AI, an ensemble-based fraud detection framework designed to enhance performance on highly imbalanced datasets while maintaining interpretability. Unlike existing ensemble approaches that rely on oversampling techniques, FraudX AI retains the original class distribution and mitigates class imbalance by optimizing model thresholds and adjusting class weights. This approach ensures that the model's predictions remain realistic in real-world banking scenarios, where fraudulent transactions are rare but highly impactful.

Our goal was to achieve higher recall compared to previous studies. FraudX AI accomplishes this by combining random forest (RF) and eXtreme Gradient Boosting (XGBoost), optimizing their predictions through threshold tuning to balance recall and precision. By prioritizing the area under the precision–recall curve (AUC-PR) and recall metrics over traditional accuracy-based metrics, this approach enables a more reliable evaluation of fraud detection models, as accuracy alone can be misleading in highly imbalanced datasets. Additionally, this study integrates SHAP to enhance model interpretability, demonstrating how SHAP values provide meaningful insights despite feature anonymization while ensuring transparency for financial analysts and regulatory compliance.

The key contributions of this research are as follows:

1. Proposing a novel fraud detection framework (FraudX AI) that enhances recall while minimizing false positives without relying on artificial data balancing techniques, ensuring real-world applicability.

2. Introducing an optimized threshold-tuning strategy to balance recall and precision, improving fraud detection effectiveness in highly imbalanced datasets.
3. Incorporating SHAP for interpretable fraud detection, providing actionable insights for financial analysts and regulatory compliance, even when working with PCA-transformed anonymized features.
4. Conducting an extensive comparative evaluation against multiple ML baselines and recent fraud detection studies, demonstrating FraudX AI's superior performance in recall and AUC-PR.
5. Discussing the generalization capability of FraudX AI and its applicability to different datasets, addressing the challenges of fraud detection across diverse financial institutions.

The remainder of this paper is structured as follows: Section 2 reviews related work on credit card fraud detection, focusing on existing methodologies and the challenges of imbalanced datasets. Section 3 presents the proposed framework and dataset, beginning with an overview of the dataset used for evaluation, followed by a detailed discussion of the FraudX AI framework and its implementation. Section 4 describes the experimental setup and results, outlining the implementation strategy, model training process, evaluation metrics, performance analysis, and SHAP-based interpretability assessment. Section 5 provides a detailed discussion of the findings, an interpretation of the results in the context of previous studies and applicability to real life, and the main limitations and challenges. Finally, Section 6 summarizes the study's contributions and suggests potential directions for future research.

2. Related Work

Fraud detection in credit card transactions has become a critical area of research due to its substantial financial implications and the increasing sophistication of fraudulent activities. The specific challenges associated with fraud detection, particularly the highly imbalanced nature of datasets and the continuously evolving fraud patterns, have led to the exploration of various ML and deep learning (DL) techniques [12–14].

A study [15] compared LR, DT, SVM, and XGBoost for fraud detection, finding XGBoost to be the best-performing model, with an accuracy of 99.88% and an AUC-ROC score of 1.0. The study demonstrated XGBoost's effectiveness in handling imbalanced datasets and applying the synthetic minority oversampling technique (SMOTE), significantly enhancing model performance. However, concerns regarding potential data skewing remained. Explainable AI tools such as SHAP and LIME enhanced XGBoost's interpretability, highlighting key fraud-related features such as transaction type and old balance. While LR and DT models attained respectable accuracies of 98.99% and 98.96%, respectively, they lacked the precision and scalability required for large-scale fraud detection. Conversely, SVM exhibited poor performance, with a recall value of only 0.36, primarily due to its reliance on linear kernels, which are inadequate for capturing complex fraud patterns. The study underscores the importance of AUC-ROC and macro-averaged metrics for evaluating fraud detection systems while acknowledging the challenges of balancing model interpretability, computational efficiency, and the scalability for real-world applications.

The authors of [16] proposed a hybrid fraud detection framework that integrated process mining and SVM, using multi-perspective features such as control flow, resource flow, and user behavior to enhance accuracy. Tools like ProM and Colored Petri nets facilitated real-time anomaly detection in e-commerce transactions, achieving a higher F1-score and AUC metrics than single-perspective models. However, the approach faced challenges, including computational overhead, scalability constraints, and difficulties adapting to evolving fraud tactics. While the model offers interpretability through process mining, its complexity may hinder practical deployment in real-world settings.

Another study [17] evaluated five ML models—ANN, SVM, RF, DT, and naive Bayes (NB)—for credit card fraud detection, with ANN achieving the highest accuracy (97.6%), followed by SVM (95.5%) and RF (94.5%). Preprocessing techniques, including data cleaning, feature scaling, and temporal analysis, were essential in improving model performance and addressing class imbalance. Feature engineering, such as incorporating transaction frequency and time-based metrics, further enhanced the discriminatory power of all models. While ANN demonstrated superior accuracy and an ability to model complex transaction patterns through backpropagation, its high computational demands were identified as a limitation. SVM also exhibited strong performance but struggled with scalability due to its computationally intensive kernel operations. DT and NB models, on the other hand, showed higher rates of false positives and false negatives, restricting their applicability in real-world fraud detection. The study underscores the necessity of balancing accuracy, recall, and precision to optimize fraud detection performance in practical applications.

Hashemi et al. [18] evaluated the performance of advanced ML models, including LightGBM, XGBoost, and CatBoost, for fraud detection. When optimized using hyperparameter tuning techniques such as Bayesian optimization, these models achieved superior performance metrics, with AUC-ROC reaching 0.95 and recall attaining 0.80. Additionally, a majority voting ensemble approach combining CatBoost, XGBoost, and LightGBM demonstrated improved results on imbalanced datasets, achieving a recall value of 0.82 and an F1-score of 0.81. Despite the success of the majority voting ensemble, the study highlighted the computational overhead associated with this approach, emphasizing the need for more efficient ensemble methods for real-world applications.

Mosa et al. [19] investigated 15 meta-heuristic optimization (MHO) algorithms for feature selection in fraud detection, identifying the sailfish optimizer (SFO) in combination with RF as particularly effective, achieving 97% accuracy while reducing the feature set size by 90%. Advanced transfer functions further improved feature selection accuracy, while undersampling effectively addressed class imbalance. Computationally efficient methods, such as particle swarm optimization (PSO) and gray wolf optimization (GWO), reduced feature dimensionality without compromising model performance. However, the study also identified key challenges, including the computational complexity of MHO techniques, the scalability limitations for large datasets, and the risks of overfitting due to overly aggressive feature selection.

The study conducted by Ming et al. [20] proposed a hybrid fraud detection framework that combines a 12-layer CNN for deep feature extraction with classifiers such as SVM, KNN, and NB. The CNN-ensemble model demonstrated high performance, achieving an AUC-ROC of 100% for credit card fraud detection. However, balanced datasets limited the applicability of these results in real-time scenarios. While ensemble methods enhanced classification performance, several limitations remained, including CNN's reduced effectiveness on smaller datasets, high computational demands, and challenges adapting to real-world, imbalanced data. Key evaluation metrics, such as AUC-ROC, precision, and recall, underscore the model's theoretical advantages but do not fully account for practical deployment constraints.

Khalid et al. [21] assessed an ensemble model incorporating SVM, KNN, RF, bagging, and boosting classifiers for credit card fraud detection, reporting high accuracy and F1-score. While RF and boosting exhibited strong performance, the training and testing times for bagging and boosting models were significantly longer, indicating inefficiencies in computational resources compared to simpler models like LR or KNN. Moreover, SMOTE-generated balanced datasets produced results that do not accurately reflect real-world fraud detection scenarios characterized by inherent class imbalance. Although SMOTE mitigates class imbalance and enhances performance, it introduces synthetic data artifacts that may

limit generalizability in operational environments. The computational inefficiencies and reliance on artificially balanced datasets highlight the challenges of deploying these models in real-time, large-scale fraud detection systems.

Another ensemble-based study [22] similarly employed SMOTE to mitigate class imbalance, proposing a weighted-average ensemble model that combined RF, KNN, bagging, adaptive boosting (AdaBoost), and LR. This approach achieved an impressive accuracy of 99.5%, while demonstrating improvements in precision, recall, and F1-score compared to individual classifiers. However, like other SMOTE-based studies, its dependence on synthetic data raises concerns regarding applicability in real-world, highly imbalanced fraud detection scenarios. Among the individual classifiers, RF and bagging exhibited the best performance, with accuracies of 98%, followed by AdaBoost and LR at 97% and KNN at 95%. The ensemble model utilized a weighted strategy to optimize predictions and reduce misclassifications. However, scalability and computational efficiency challenges persist, particularly in real-time fraud detection applications. These findings underscore the broader limitations of SMOTE-based ensemble methods in operational environments.

Several recent studies have explored advanced ML and DL techniques for fraud detection. Notable bodies of research by Tian et al. [23] and Xie et al. [24–26] have contributed significantly to developing graph-based and sequence-based fraud detection models. Graph neural networks (GNNs) have gained popularity in transactional fraud detection due to their ability to capture transaction relationships. In [23], ASA-GNN was introduced as an adaptive sampling and aggregation approach to improve fraud classification performance by filtering noisy nodes and capturing behavioral patterns through graph-based modeling.

Xie et al. [24] have developed multiple DL-based fraud detection models, emphasizing behavioral pattern extraction from sequential transaction data. The authors introduced the time-aware interaction LSTM (TAI-LSTM) model, designed to extract transactional behaviors and learn transactional behavioral representations for each transaction record within a user's history. Their approach enables the model to enhance fraud detection by learning both short-term and long-term transactional habits. Similarly, their spatial–temporal gated network (STGN) introduces spatio-temporal attention mechanisms to model credit card transactions dynamically, capturing fraudulent transactions' geographical and temporal distribution [25]. By incorporating spatial and temporal features, STGN improves fraud detection by identifying the transaction location and timing inconsistencies. The authors also expanded their research in [26] by introducing the time-aware historical-attention-based LSTM (TH-LSTM), a model designed to capture transactional behavioral representations by leveraging time-aware LSTM networks. This approach effectively extracts long-term dependencies in transaction sequences, enabling the model to identify subtle spending patterns and detect fraudulent activities more accurately.

Table 1 presents a comparative analysis of recent studies, examining their methodologies, datasets, results, and limitations. This analysis highlights advancements in the field and persistent challenges, including scalability, reliance on synthetic data, and computational efficiency in real-world applications.

While the reviewed studies demonstrate significant advancements in credit card fraud detection through ML, DL, and ensemble methods, several critical challenges remain. The reliance on SMOTE and balanced datasets, although effective in improving performance metrics, limits the generalizability of these models to real-world scenarios where data are highly imbalanced. Moreover, computational inefficiencies and scalability issues pose obstacles to deploying these methods in real-time, high-volume environments. Additionally, many models' "black-box" nature raises concerns regarding interpretability.

Table 1. Analysis of fraud detection studies.

Reference	Detection Approaches	Dataset	Preprocessing	Metrics/ Results	Limitations
Nobel et al. [15]	LR, DT, SVM, XGBoost	A synthetic PaySim dataset	Balanced with SMOTE	XGBoost: accuracy 99.88%, precision 96%, recall 88%; SVM: accuracy 96.91%, precision 64%, recall 36%	Results on balanced datasets; low recall limits real-world applicability
Yu et al. [16]	Hybrid framework: Process mining + SVM	Event logs	Multi-perspective	Precision 95%, recall 85%, F1-score 90%, AUC-ROC 94%	Computational overhead and scalability challenges; need to adapt to evolving fraud tactics
Jayanthi et al. [17]	ANN, SVM, RF, DT, NB	Credit card fraud dataset	Oversampling (fraudulent transactions) and undersampling (legitimate transactions), feature engineering, temporal considerations	ANN: accuracy 98%; SVM: accuracy 95.5%; RF: accuracy 94.5%	High computational demands of ANN; oversampling and undersampling create an unrealistic dataset for real-world applications
Hashemi et al. [18]	LightGBM, XGBoost, CatBoost	European credit card transactions	Unbalanced datasets, hyperparameter tuning	Combining the LightGBM and XGBoost methods: accuracy 99.93%, precision 79%, recall 80%, AUC-ROC 95.2%, F1-score 79%	High computational overhead and limited efficiency of majority voting ensemble
Mosa et al. [19]	A total of 15 MHO algorithms (SFO, PSO, GWO) + RF	European credit card transactions	Undersampling, advanced transfer functions	SFO + RF: accuracy 97% (on the imbalanced test set)	High computational complexity of MHO techniques; scalability challenges and reliance solely on accuracy metrics
Ming et al. [20]	Twelve-layer CNN + SVM, KNN, NB	European credit card transactions	Balanced with oversampling techniques	CNN with ensemble ML models: accuracy 99.99%, precision 100%, recall 99.97%, F1-score 100%, AUC-ROC 100%	Computational demands and dataset balance requirements constrain real-time applicability
Khalid et al. [21]	SVM, KNN, RF, Bagging, Boosting	European credit card transactions	Balanced with SMOTE	PM (SMOTE): accuracy 99.96%, precision 99.96%, recall 99.96%, F1-score 99.96%	Computational inefficiency and evaluation on balanced datasets make the approach unsuitable for real-world imbalanced scenarios
Sahithi et al. [22]	Weighted-Average Ensemble (RF, KNN, Bagging, AdaBoost, LR)	European credit card transactions	Balanced with SMOTE	PM: accuracy 99.94%, precision 99.95%, recall 99.94%, F1-score 99.95%	Computational inefficiency and evaluation on balanced datasets make the approach unsuitable for real-world imbalanced scenarios

Table 1. *Cont.*

Reference	Detection Approaches	Dataset	Preprocessing	Metrics/ Results	Limitations
Tian et al. [23]	ASA-GNN	Private, TC, and XF datasets	Graph-based aggregation of transaction relationships	Private: recall 88.4%, F1-score 90.4%, AUC 92.2%; XF: recall 73.1%, F1-score 59.3%, AUC 57.1%; TC: recall 86.8%, F1-score 89.2%, AUC 88.4%	Computationally intensive; requires well-defined graph structures
Xie et al. [24]	TAI-LSTM	Large real-world dataset and European credit card transactions	Time-aware feature extraction and sequence modeling	Real-world dataset (average value): precision 88.3%, recall 93.8%, F1-score 90.5%, AUC 95.2%; European credit card transactions: precision 50.7%, recall 99.6%, F1-score 67.2%, AUC 51%	Reliance on transaction timestamps and sequential patterns, making it less effective for new users
Xie et al. [25]	STGN	Real-world transaction dataset	Spatio-temporal feature extraction and representation learning	Average value: precision 89%, recall 94.6%, F1-score 91.3%, AUC 96.6%	High complexity due to spatio-temporal modeling; cold-start problem
Xie et al. [26]	TH-LSTM	Large-scale real-world transaction dataset and European credit card transactions	Transaction sequence representation with behavioral learning	Real-world dataset (average value): precision 88%, recall 90.8%, F1-score 88.1%, AUC 94.4%; European credit card transactions: precision 50.4%, recall 99.6%, F1-score 66.9%, AUC 50.4%	Heavy reliance on labeled data; struggles with generalization to unseen fraud patterns

To address these challenges, our study proposes an alternative approach that eliminates the need for synthetic data augmentation while maintaining high recall and precision. Unlike prior works, we leverage class weight adjustments instead of SMOTE, ensuring that the model learns directly from the original data distribution. Furthermore, during training, we introduce a multi-layer perceptron (MLP) validator, enhancing feature extraction without affecting inference time predictions. Our method achieves superior recall and AUC-PR scores and enhances model transparency through SHAP-based interpretability, making it well suited for real-world financial applications.

3. Proposed Framework and Dataset

This section introduces the dataset used to evaluate the FraudX AI framework, outlining its characteristics and challenges and a detailed discussion of its design and implementation.

3.1. Dataset

We assessed the effectiveness of FraudX AI using the European credit card fraud dataset [27], widely recognized as a benchmark in fraud detection research. Its accessibility and extensive use in prior studies make it a valuable resource for comparative analysis and situating findings within the broader context of current research in this

field. The dataset contains 284,807 transactions recorded over two days, among which only 492 transactions (0.172%) are fraudulent (Figure 1). This pronounced class imbalance closely reflects real-world conditions, underscoring the inherent challenges of fraud detection.

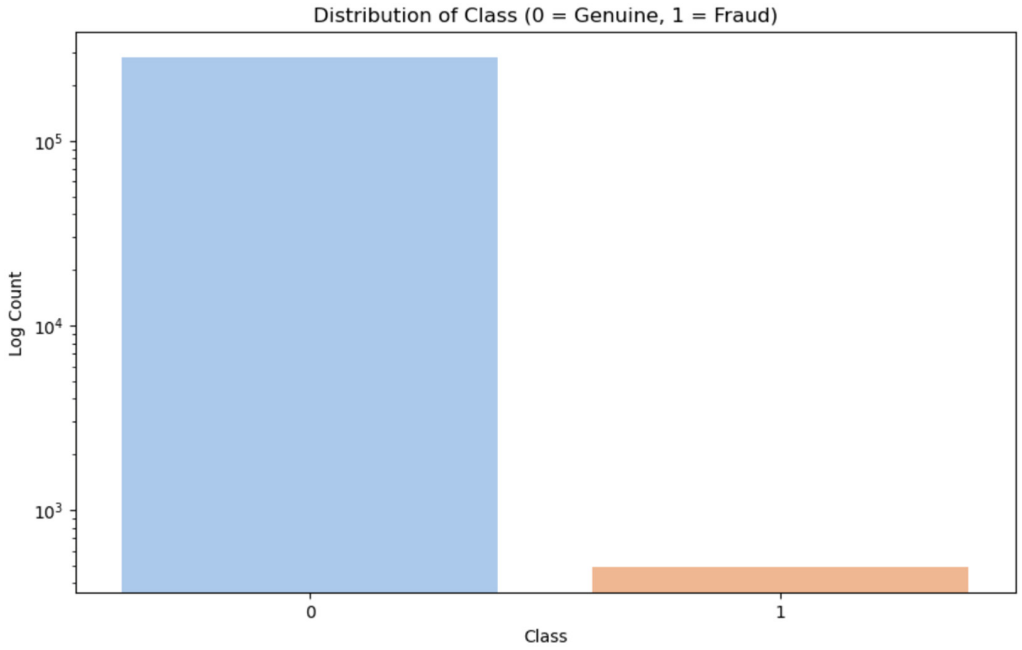

Figure 1. Distribution of class labels in the dataset.

Each transaction comprises 31 features, including 28 anonymized numerical variables (V1–V28) transformed using Principal Component Analysis (PCA) to preserve confidentiality. The remaining features were retained in their original form, as follows: Time, representing the elapsed time between transactions; Amount, indicating the monetary value of each transaction; and Class, which serves as the target variable, where a value of 0 denotes a genuine transaction and 1 represents fraud.

Due to the severe class imbalance, traditional ML models trained directly on this dataset tend to favor the majority class, resulting in high overall accuracy but poor recall for fraud detection. This study evaluates fraud detection models on the original imbalanced dataset to ensure real-world applicability, ensuring that performance metrics accurately reflect operational challenges.

We divided the dataset into 80% for training and 20% for testing to ensure model evaluation while maintaining class distribution. Additionally, we implemented feature scaling techniques to normalize transaction values and improve model convergence.

Given the complexity of fraud detection in highly imbalanced datasets, a specialized ensemble-based framework was required to improve fraud identification while maintaining interpretability. The following subsection introduces FraudX AI, outlining its design, components, and methodological approach.

3.2. Overview of FraudX AI Framework

FraudX AI is an ensemble-based fraud detection framework designed to maximize recall and AUC-PR while ensuring interpretability. The architecture, illustrated in Figure 2, outlines the workflow from dataset preprocessing to model evaluation and explainability.

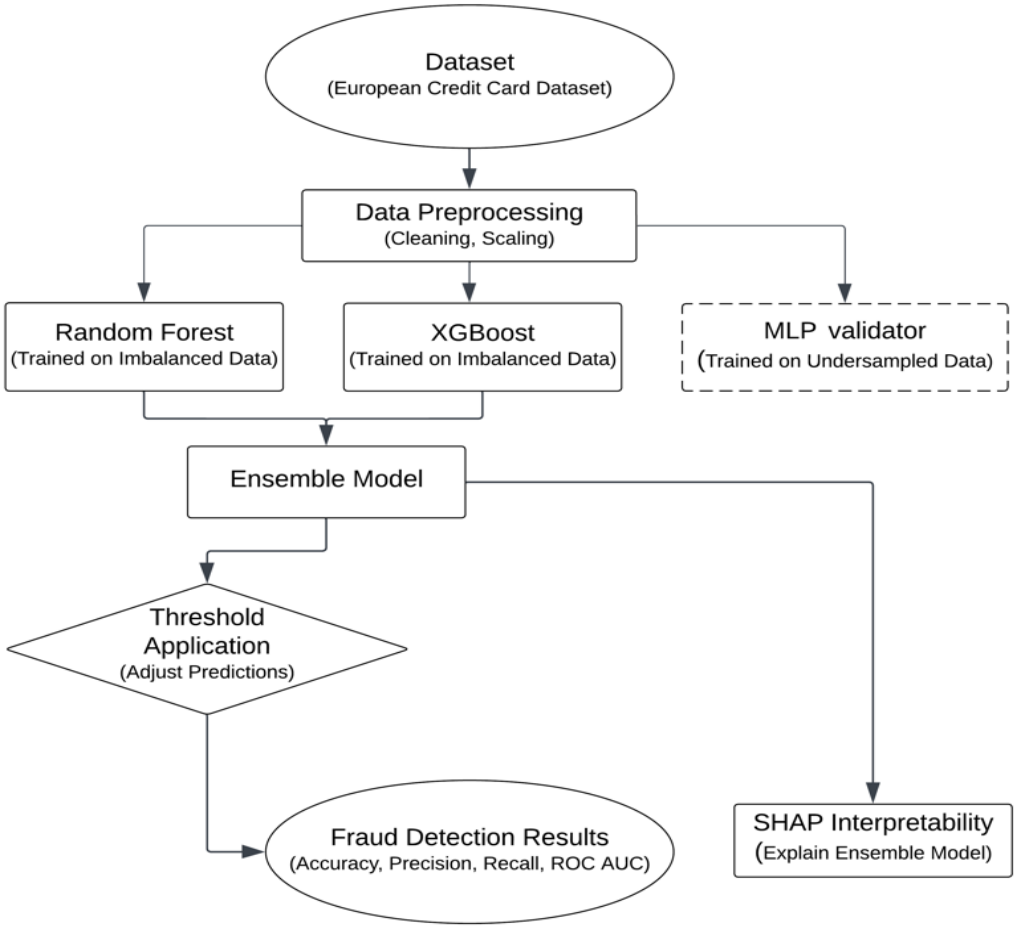

Figure 2. The architecture of the FraudX AI framework.

As shown in Figure 2, the proposed framework comprises several key stages, as follows: data preprocessing, model training, ensemble learning, threshold optimization, and SHAP-based interpretability.

The first stage, data preprocessing, involves cleaning, scaling, and preparing raw transaction data for training. Given the highly imbalanced nature of fraud detection datasets, an MLP validator trained on an undersampled dataset is incorporated as a reference model to assess the impact of class balancing. However, this validator does not directly contribute to the final fraud detection decisions.

The core fraud detection pipeline consists of two base models, RF and XGBoost, trained on the original imbalanced dataset to preserve real-world class distributions. These models independently generate probability scores for fraudulent transactions. Their outputs are then combined using probability-weighted averaging, leveraging the complementary strengths of both classifiers to enhance predictive performance.

To refine fraud detection outcomes, threshold optimization is applied to balance precision and recall, ensuring a high fraud detection rate while minimizing false positives. This step is particularly critical in fraud detection systems, where undetected fraudulent transactions can lead to significant financial losses.

Finally, SHAP enhances interpretability by quantifying how individual features contribute to the model's predictions. This approach improves transparency, enabling stakeholders to understand the reasoning behind fraud classifications, which is essential for regulatory compliance and decision making. However, since PCA was applied to transform the original feature space and ensure data privacy in the European credit card fraud dataset, the actual feature names are unavailable, with all features being represented as anonymized

components (V1 to V28). While this may reduce interpretability, we acknowledge that individuals with access to the original feature mappings can derive more precise insights into the model's decision-making process.

4. Experimental Setup and Results

4.1. Implementation and Training Strategy

The FraudX AI framework consists of two base models (RF and XGBoost) trained on the original imbalanced dataset. In contrast, an MLP validator was trained on an undersampled subset to assess preprocessing effectiveness. Notably, the MLP validator does not directly contribute to ensemble decision making.

We configured the models with standard hyperparameter settings, optimizing them through validation experiments. The RF and XGBoost models were trained using balanced learning strategies, incorporating class weighting to address class imbalance effectively. Specifically, a class weight ratio of {0:1.0, 1:2.0} was applied to assign higher importance to the minority fraud class during training.

The MLP validator employs a fully connected neural network architecture with a single hidden layer, leveraging gradient-based optimization for training. To maximize model performance, hyperparameters such as the number of trees, learning rates, and activation functions were fine-tuned through empirical analysis.

FraudX AI utilizes an ensemble learning approach, aggregating RF and XGBoost predictions through a probability-weighted strategy. Model probabilities are combined using an empirically determined weighting scheme, leveraging the strengths of both classifiers.

Since fraud detection requires balancing recall (capturing fraudulent cases) and precision (reducing false positives), adaptive threshold tuning was applied instead of a fixed classification threshold. The optimal threshold was determined through precision–recall curve analysis, ensuring a high fraud detection rate while minimizing false alarms.

All experiments were conducted on a MacBook Pro equipped with an Apple M1 Pro chip, featuring a multi-core CPU and integrated GPU acceleration, which provided sufficient computational power for data preprocessing, model training, and evaluation. The implementation was carried out in a Jupyter Notebook 7.3.2 environment, facilitating efficient development, testing, and visualization.

We implemented the framework using standard ML and DL libraries, including scikit-learn, XGBoost, TensorFlow/Keras, pandas, NumPy, and visualization tools like Matplotlib 3.9.2 and Seaborn 0.13.2. Training efficiency was optimized to ensure model convergence within a reasonable computational timeframe.

4.2. Evaluation Metrics

Selecting the appropriate metric is essential for evaluating the performance of ML models. Since fraud is a rare event, evaluation metrics must accurately reflect a model's ability to handle class imbalance. Many studies emphasize accuracy as a primary metric. Still, in highly imbalanced datasets where most transactions are genuine, accuracy can be misleading, as a model that classifies all transactions as non-fraudulent may achieve high accuracy while failing to detect actual fraud cases, rendering it ineffective in real-world applications [28,29]. To address this challenge, more reliable evaluation metrics, such as recall (sensitivity) and AUC-PR, are preferred [30]. Recall measures the system's ability to identify fraudulent transactions, which is critical for minimizing the risk of undetected fraud. In contrast to traditional AUC-ROC, AUC-PR focuses on the minority class, making it more suitable for fraud detection, as it better reflects performance in imbalanced settings [31].

Key performance metrics, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), are used to assess the model's effectiveness. These values are the foundation for computing accuracy, precision, recall, F1-score, and AUC-PR. The following equations define these metrics mathematically, comprehensively evaluating fraud detection models.

Accuracy is the proportion of correctly predicted cases (both positive and negative) out of the total number of cases [32]. A high accuracy indicates that the model correctly predicts most cases, and, therefore, it can be misleading in imbalanced datasets. For instance, in a dataset where 99% of transactions are genuine (non-fraudulent), a model predicting all transactions as genuine will have 99% accuracy despite being ineffective at detecting fraud. Equations (1)–(4) present the mathematical formulations for accuracy, precision, recall, F1-score, and AUC-PR.

$$\text{Accuracy} = \text{TN} + \text{TP}/\text{TN} + \text{FP} + \text{TP} + \text{FN} \tag{1}$$

Precision is a performance metric that quantifies the proportion of correctly predicted positive cases (fraud) out of all cases classified as positive by the model [33]. It measures the model's ability to avoid false positives (false alarms). In fraud detection, a model with low precision may misclassify many legitimate transactions as fraudulent, leading to unnecessary inconvenience for users. High precision reduces operational costs and maintains user trust by minimizing false alarms.

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{2}$$

Recall, also called sensitivity or true positive rate (TPR), evaluates a model's ability to correctly identify all actual positive cases in a dataset [34]. In fraud detection, recall indicates how effectively the model identifies fraudulent transactions among all actual fraud cases. A high recall value signifies that the model successfully detects most fraudulent transactions, minimizing false negatives. Conversely, a low recall value suggests that the model fails to detect a significant number of fraud cases, which can have severe financial and reputational consequences.

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{3}$$

The F1-score is a performance metric that balances precision and recall [35]. It is beneficial for evaluating models in scenarios with class imbalance.

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \tag{4}$$

Since our framework incorporates a threshold optimization step to balance recall and precision, evaluating performance using the AUC-PR metric was essential. AUC-PR measures a model's ability to maintain an optimal trade-off between precision and recall across all classification thresholds. It is a single scalar value that summarizes the precision–recall (PR) curve, and it is computed as the precision weighted by the probability that the output score Y falls below a given threshold c [36]. Mathematically, it is defined as follows:

$$\text{AUC} - \text{PR} = \int_{-\infty}^{\infty} \text{Prec}(c)\text{dP}(Y \leq c) \tag{5}$$

where Prec(c) is the precision at threshold c and $P(Y \leq c)$ is the probability distribution of the output scores.

The AUC-PR value ranges between 0 and 1, reflecting the model's ability to balance precision and recall across varying thresholds. In imbalanced datasets, AUC-PR

is particularly advantageous because it avoids the pitfalls of misleading accuracy metrics and provides a more robust measure of model performance for rare events such as fraud detection [37].

4.3. Results and Analysis

To evaluate the performance of FraudX AI, we compared it with the following eight baseline ML models: LR, DT, RF, gradient boosting (GB), XGBoost, LightGBM, NB, and MLP. All models were trained and tested on the imbalanced dataset, adhering to the previously defined data split. The performance was assessed using the evaluation metrics outlined earlier, as presented in Tables 2 and 3.

Table 2. Performance metrics for the testing dataset.

Models	Accuracy	Precision	Recall	F1-Score	AUC-PR
Random Forest	0.99	0.96	0.74	0.84	0.86
Logistic Regression	0.97	0.05	0.92	0.1	0.74
Decision Tree	0.99	0.68	0.72	0.7	0.7
XGBoost	0.99	0.92	0.81	0.86	0.88
MLP	0.99	0.65	0.60	0.63	0.54
Gradient Boosting	0.99	0.53	0.18	0.27	0.21
LightGBM	0.99	0.82	0.87	0.84	0.89
Naive Bayes	0.99	0.14	0.66	0.23	0.35
FraudX AI	0.99	1.00	0.95	0.97	0.97

Table 3. Confusion matrix values for the testing dataset.

Models	TN	FP	FN	TP
Random Forest	56,861	3	25	73
Logistic Regression	55,234	1630	8	90
Decision Tree	56,830	34	27	71
XGBoost	56,857	7	19	79
MLP	56,833	31	39	59
Gradient Boosting	56,848	16	80	18
LightGBM	56,845	19	13	85
Naive Bayes	56,457	407	33	65
FraudX AI	56,864	0	5	93

While Tables 2 and 3 provide a detailed numerical comparison of the performance metrics for all tested models, Figure 3 visualizes the results specifically in terms of recall and AUC-PR. This visualization highlights the models' ability to identify fraudulent transactions (recall) and their effectiveness in distinguishing between fraudulent and legitimate transactions (AUC-PR). This graphical representation clearly illustrates the differences in performance, offering a more intuitive understanding of how FraudX AI surpasses baseline models.

To further validate the effectiveness of FraudX AI compared to baseline models, we evaluated our proposed framework against previous studies that used the same unbalanced credit card fraud dataset [27]. Table 4 provides a comparative analysis, demonstrating how our framework outperforms prior research in key evaluation metrics, such as achieving a balance between high recall, high precision, and AUC-PR. This comparison highlights the robustness of our framework relative to traditional individual-based models and state-of-the-art methods, ensuring its competitiveness and applicability in real-world scenarios.

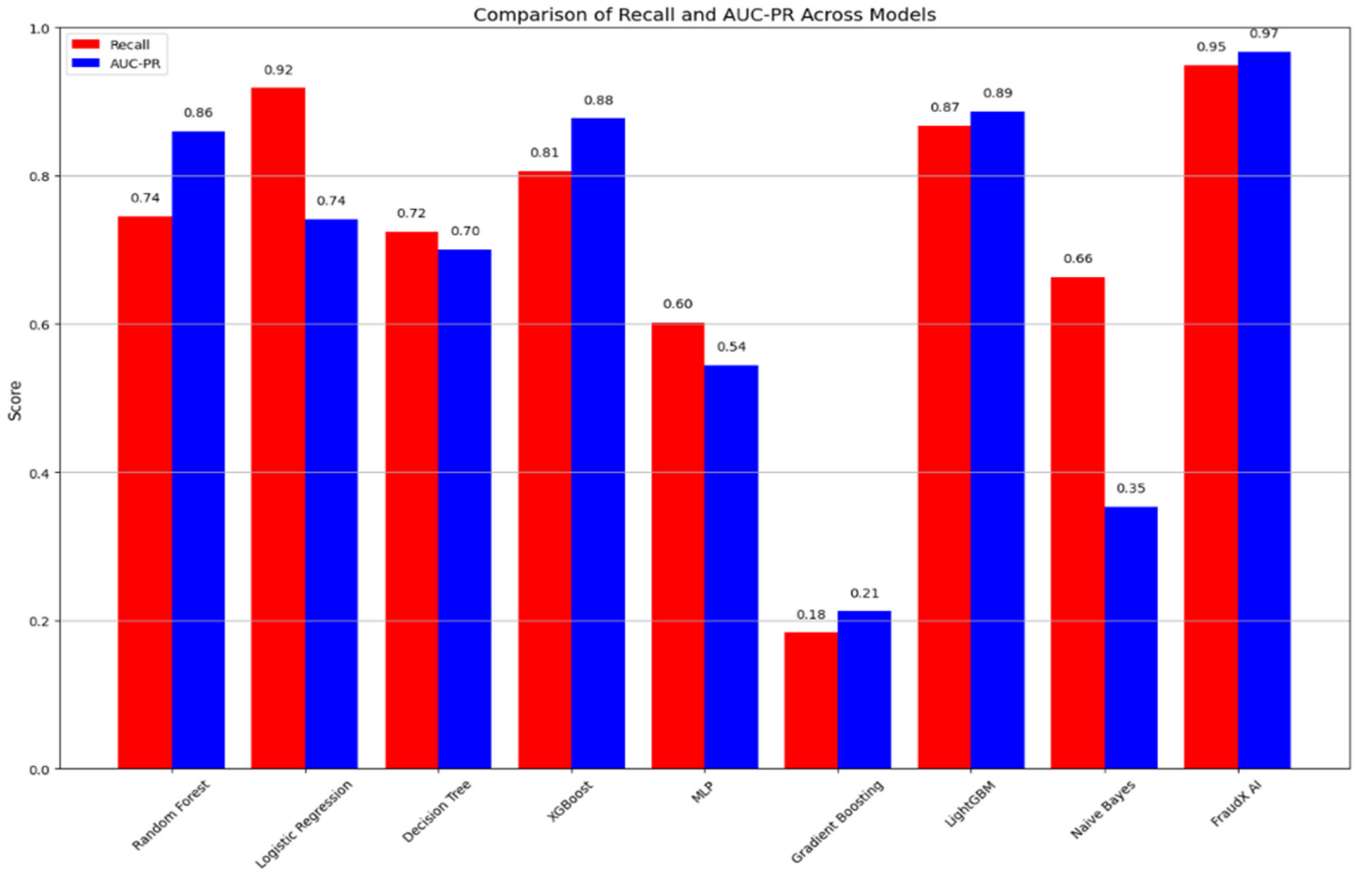

Figure 3. Comparison of recall and AUC-PR across fraud detection models.

Table 4. Comparison of the proposed framework with existing research.

Methods/Studies	Accuracy	Precision	Recall	F1-Score	AUC-PR
TAI-LSTM [24]	-	51%	99%	67%	-
TH-LSTM [26]	-	50%	99%	67%	-
Weighted DNN [31]	98%	7%	89%	13%	-
FMC [38]	96%	94%	91%	92%	-
CB-CHL-LGBM [39]	-	74%	84%	-	-
ANN [40]	93%	97%	90%	-	-
Multiple Classifier [41]	99%	-	87%	-	-
FraudX AI	99%	100%	95%	97%	97%

For instance, the TAI-LSTM and TH-LSTM models demonstrate impressive capabilities, with a remarkable recall of 99%. However, these models also show a precision score of 51% and 50%, respectivley, resulting in an F1-score of 67% for both. While these results highlight their ability to detect a high proportion of fraud cases, they also indicate a trade-off with precision, where a significant number of flagged transactions may be legitimate. Methods such as FMC and the weighted DNN achieved recall values of 91% and 89%, but their precision scores were either lower or inconsistent. Similarly, the CB-CHL-LGBM method achieved a precision score of 74%, but its recall was limited to 84%, indicating challenges in balancing these key metrics. FraudX AI effectively addresses these limitations by achieving well-balanced precision and recall of 100% and 95%, respectively, leading to a higher F1-score (97%) and AUC-PR (97%). This balance between precision and recall is critical for fraud detection systems, as it minimizes false positives while maximizing the identification of fraudulent transactions.

Additionally, FraudX AI outperforms traditional models like ANN and the multiple classifier approach, which, while effective, fail to achieve comparable results across all metrics. By integrating advanced ML techniques with threshold optimization and explainability features, FraudX AI surpasses existing approaches and ensures transparency and reliability, making it a competitive solution for deployment in real-world industry settings.

4.4. SHAP Analysis for FraudX AI Interpretability

Model interpretability is a cornerstone of trust and transparency in credit card fraud detection systems. Integrating SHAP into our proposed framework provides a comprehensive understanding of how various features influence the model's decisions. This aspect is crucial for industry stakeholders and regulatory compliance, where explainability is as vital as the model's ability to detect fraudulent transactions effectively. By analyzing SHAP values, we gain insight into feature contributions, ensuring that the decision-making process is interpretable and compliant with financial regulations.

In this study, we utilized SHAP summary plots to visualize the overall impact of features on fraud classification and SHAP-based importance analysis to assess the influence of specific features in fraudulent transactions. Figure 4 shows the SHAP summary plot, ranking features according to their contribution to fraud classification. The results indicate that V4, V14, and V12 significantly impact the model decisions, suggesting that these transaction attributes play a key role in distinguishing between fraudulent and genuine transactions. Notably, features such as V4 and V11 exhibit a wide range of SHAP values, implying that high and low values influence the classification outcomes. In contrast, V14 and V12 display a more polarized effect, indicating a more substantial directional influence on fraud detection.

Figure 4. SHAP summary plot of feature importance in FraudX AI.

While Figure 4 presents the overall feature importance in FraudX AI, Figure 5 refines this analysis further by illustrating feature importance specifically for fraudulent transactions. This visualization highlights the key features that most influence classifying fraudulent cases, offering a more targeted approach to fraud detection.

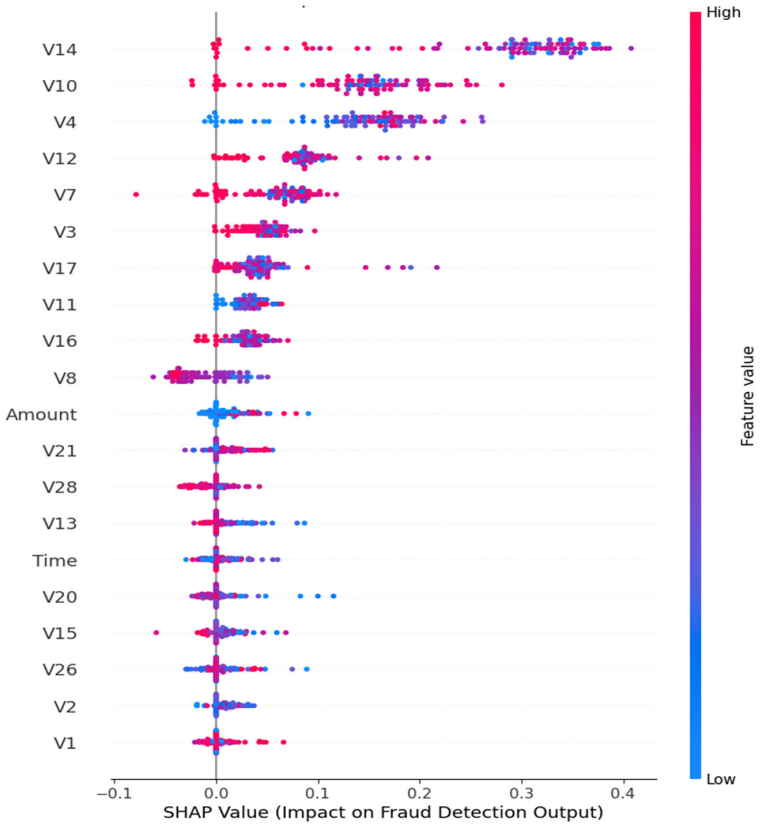

Figure 5. SHAP feature importance for fraudulent transactions in FraudX AI.

As shown in Figure 5, V14, V10, and V4 emerge as the most influential features in detecting fraudulent transactions, exhibiting high SHAP values. It indicates that variations in these features significantly impact the model's decision to classify a transaction as fraudulent. Interestingly, V12, V7, and V3 also play a crucial role, reinforcing that fraud detection relies on multiple interacting factors rather than on a single dominant variable.

The Amount feature, which represents the transaction value, exhibits lower importance than some of the transformed PCA features. It aligns with previous findings showing that fraudsters often manipulate the transaction value to evade detection, making more complex behavioral features crucial for classification. Additionally, the Time feature demonstrates moderate importance, indicating that transaction timing may contribute to anomaly detection.

By comparing Figures 4 and 5, we observe that, while the overall feature significance remains relatively consistent, certain variables exert a more decisive impact, specifically in fraudulent cases. This difference underscores the value of SHAP in providing a granular interpretation, enabling fraud analysts to focus on the most critical attributes when investigating potential fraudulent activity.

The SHAP analysis highlights key features influencing the FraudX AI decision-making process, offering a transparent view of fraud classification. By identifying the most influential variables, this interpretability enhances confidence in the model's predictions and supports financial institutions in making informed risk management decisions. Furthermore, this explainability strengthens the applicability of FraudX AI in real-world fraud detection scenarios, where the accuracy and validity of model decisions are essential.

5. Discussion

The FraudX AI framework demonstrates promising results for credit card fraud detection, achieving the highest recall of 95% on an unbalanced dataset while maintaining an AUC-PR of 97%. These results underscore the effectiveness of the ensemble approach combining RF and XGBoost, optimized through MLP validation, class weighting, threshold tuning, and probability averaging.

A primary challenge in credit card fraud detection is the imbalanced classification problem, where fraudulent transactions constitute a small fraction of all transactions. Models trained on imbalanced datasets often prioritize high accuracy, leading to misleading results where all transactions are classified as legitimate to achieve high accuracy without effectively detecting fraud.

FraudX AI introduces an innovative approach to address these challenges without relying on traditional oversampling techniques like SMOTE. By retaining the original imbalanced dataset and incorporating advanced techniques such as MLP validation during training, class-weighted learning, and threshold tuning, FraudX AI enhances model robustness and ensures accurate fraud detection outcomes while preserving real-world transaction distributions.

Moreover, while SHAP has been utilized in fraud detection, previous studies primarily focused on DL models with domain-specific feature engineering. FraudX AI integrates SHAP with ensemble learning, balancing interpretability and performance. However, the interpretation of features in this study is limited due to the PCA-transformed dataset. Future research will explore datasets with non-anonymized features to better correlate SHAP results with real-world transactional attributes. Despite this limitation, the model's explainability is not solely based on SHAP, but also on its structured threshold optimization and class-weighted learning approach, which enables financial institutions to interpret predictions without relying on opaque DL architectures.

All comparisons in this study were exclusively conducted with research utilizing the same European credit card fraud dataset and avoiding oversampling or undersampling techniques. It ensures a fair evaluation of FraudX AI's performance compared to similar methodologies, maintaining the integrity of the original class distribution and validating the model's effectiveness in practical fraud detection scenarios.

Future work will expand the evaluation of FraudX AI to include diverse financial datasets, including the one developed in previous work [42], to validate its adaptability and effectiveness in detecting evolving fraud behaviors across various transaction environments. These efforts aim to enhance the model's generalizability across different fraud patterns and transaction environments, improving its adaptability and effectiveness in real-world applications.

6. Conclusions

In conclusion, FraudX AI significantly advances fraud detection methodologies for credit card transactions. The framework achieves high recall and precision rates, demonstrating its effectiveness in identifying fraudulent transactions while minimizing false positives.

FraudX AI's approach to handling imbalanced datasets and integrating advanced methodologies such as class-weighted learning, MLP validation, and SHAP sets it apart from previous fraud detection ensemble models. FraudX AI optimizes fraud detection while preserving model integrity, providing financial institutions with a scalable, interpretable solution for mitigating fraud risks. The methodological adaptability of the framework enhances fraud detection strategies in real-world financial ecosystems.

Future research will prioritize validating FraudX AI's applicability across diverse datasets and implementing adaptive learning mechanisms to refine fraud detection strategies in response to emerging fraud trends. Additionally, there will be an investigation into extending FraudX AI's evaluation beyond credit card fraud detection to encompass other domains, such as network anomaly detection, where anomalies are infrequent events. This initiative aims to rigorously validate the model's robustness and effectiveness in identifying various types of anomalies across a wide range of operational environments.

In summary, FraudX AI sets new standards in fraud detection methodologies, promising enhanced recall, interpretability, and adaptability crucial for addressing the evolving challenges of fraud in modern financial environments.

Author Contributions: Conceptualization, N.B., S.G. and J.E.D.; methodology, N.B.; software, N.B.; validation, N.B., S.G., M.T. and J.E.D.; formal analysis, N.B. and M.T.; investigation, N.B.; resources, E.T.M. and K.B.; data curation, N.B.; writing—original draft preparation, N.B.; writing—review and editing, N.B., J.E.D. and E.T.M.; visualization, N.B.; supervision, S.G., J.E.D. and M.T.; project administration, M.T. and S.G.; funding acquisition, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP25794007).

Data Availability Statement: The dataset used in this study, the Credit Card Fraud Detection Dataset, was provided by Worldline and the Machine Learning Group of ULB. It is publicly available on Kaggle at the following link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (accessed on 10 December 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
ANNs	Artificial Neural Networks
LR	Logistic Regression
DT	Decision Tree
SVMs	Support Vector Machines
SHAP	Shapley Additive Explanations
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
AUC-PR	Area Under the Precision–Recall Curve
DL	Deep Learning
AUC-ROC	Area Under the ROC Curve
SMOTE	Synthetic Minority Oversampling Technique
LIME	Local Interpretable Model-Agnostic Explanations
NB	Naive Bayes
LightGBM	Light Gradient-Boosting Machine
CatBoost	Category Boosting
MHO	Meta-Heuristic Optimization
SFO	Sailfish Optimizer
PSO	Particle Swarm Optimization
GWO	Gray Wolf Optimization
SMOTE	Synthetic Minority Oversampling Technique
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbors

AdaBoost	Adaptive Boosting
GNNs	Graph Neural Networks
TAI-LSTM	Time-Aware Interaction-Long Short-Term Memory
STGN	Spatial–Temporal Gated Network
TH-LSTM	Time-Aware Historical-Attention-Based LSTM
MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
GB	Gradient Boosting

References

1. Prajapati, D.; Tripathi, A.; Mehta, J.; Jhaveri, K.; Kelkar, V. Credit Card Fraud Detection Using Machine Learning. In Proceedings of the 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), Mumbai, India, 3–4 December 2021; pp. 1–6. [CrossRef]
2. Juniper Research. Online Payment Fraud: Market Forecasts, Emerging Threats & Segment Analysis 2023–2028, Juniper Research. 2023. Available online: https://www.juniperresearch.com/research/fintech-payments/fraud-identity/online-payment-fraud-research-report/ (accessed on 9 January 2025).
3. Carcillo, F.; Le Borgne, Y.A.; Caelen, O.; Kessaci, Y.; Oblé, F.; Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. *J. Inf. Sci.* **2019**, *557*, 317–331. [CrossRef]
4. Makolo, A.; Adeboye, T. Credit Card Fraud Detection System Using Machine Learning. *Int. J. Inf. Technol. Comput. Sci. (IJITCS)* **2021**, *13*, 24–37. [CrossRef]
5. Yousefimehr, B.; Ghatee, M. A distribution-preserving method for resampling combined with LightGBM-LSTM for sequence-wise fraud detection in credit card transactions. *Expert Syst. Appl.* **2025**, *262*, 125661. [CrossRef]
6. Bahadur, S.; Jha, S.K. Analysis of machine learning and deep learning techniques for credit card fraud detection in class imbalanced datasets. In *Computational Methods in Science and Technology: Proceedings of the 4th International Conference on Computational Methods in Science & Technology (ICCMST 2024), Mohali, India, 2–3 May 2024*, 1st ed.; Kaur, S., Kamboj, S., Kumar, M., Dagur, A., Shukla, D.K., Eds.; CRC Press: Boca Raton, FL, USA, 2024; Volume 1, pp. 115–122. [CrossRef]
7. Thanathamathee, P.; Sawangarreerak, S.; Chantamunee, S.; Nizam, D.N.M. SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection. *Emerg. Sci. J.* **2024**, *8*, 2404–2430. [CrossRef]
8. Wang, Z.; Chen, X.; Wu, Y.; Jiang, L.; Lin, S.; Qiu, G. A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud. *Sci. Rep.* **2025**, *15*, 218. [CrossRef]
9. Zhang, L.; Xuan, Y.; Liu, Z.; Du, Z.; Wang, S.; Wang, J. A hybrid ensemble model to detect Bitcoin fraudulent transactions. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109810. [CrossRef]
10. Vasconcelos, M.; Cavique, L. Mitigating false negatives in imbalanced datasets: An ensemble approach. *Expert Syst. Appl.* **2025**, *262*, 125674. [CrossRef]
11. Zeng, Q.; Lin, L.; Jiang, R.; Huang, W.; Lin, D. NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks. *Inf. Process. Manag.* **2025**, *62*, 103916. [CrossRef]
12. Gupta, P.; Varshney, A.; Khan, M.R.; Ahmed, R.; Shuaib, M.; Alam, S. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. *Procedia Comput. Sci.* **2023**, *218*, 2575–2584. [CrossRef]
13. Cherif, A.; Badhib, A.; Ammar, H.; Alshehri, S.; Kalkatawi, M.; Imine, A. Credit card fraud detection in the era of disruptive technologies: A systematic review. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 145–174. [CrossRef]
14. Mienye, I.D.; Jere, N. Deep Learning for Credit Card Fraud Detection: A Review of Algorithms, Challenges, and Solutions. *IEEE Access* **2024**, *12*, 96893–96910. [CrossRef]
15. Nobel, S.M.N.; Sultana, S.; Singha, S.P.; Chaki, S.; Mahi, M.J.N.; Jan, T.; Barros, A.; Whaiduzzaman, M. Unmasking Banking Fraud: Unleashing the Power of Machine Learning and Explainable AI (XAI) on Imbalanced Data. *Information* **2024**, *15*, 298. [CrossRef]
16. Yu, W.; Wang, Y.; Liu, L.; An, Y.; Yuan, B.; Panneerselvam, J. A Multiperspective Fraud Detection Method for Multiparticipant E-Comerce Transactions. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 1564–1576. [CrossRef]
17. Jayanthi, G.; Deepthi, P.; Rao, B.N.; Bharathiraja, M.; Logapriya, A. A Comparative Study on Machine Learning and Fuzzy Logic-Based Approach for Enhancing Credit Card Fraud Detection. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 192–199.
18. Hashemi, S.K.; Mirtaheri, S.L.; Greco, S. Fraud Detection in Banking Data by Machine Learning Techniques. *IEEE Access* **2023**, *11*, 3034–3043. [CrossRef]

19. Mosa, D.T.; Sorour, S.E.; Abohany, A.A.; Maghraby, F.A. CCFD: Efficient Credit Card Fraud Detection Using Meta-Heuristic Techniques and Machine Learning Algorithms. *Mathematics* **2024**, *12*, 2250. [CrossRef]
20. Ming, R.; Abdelrahman, O.; Innab, N.; Ibrahim, M.H.K. Enhancing fraud detection in auto insurance and credit card transactions: A novel approach integrating CNNs and machine learning algorithms. *PeerJ Comput. Sci.* **2024**, *10*, e2088. [CrossRef]
21. Khalid, A.R.; Owoh, N.; Uthmani, O.; Ashawa, M.; Osamor, J.; Adejoh, J. Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach. *Big Data Cogn. Comput* **2024**, *8*, 6. [CrossRef]
22. Sahithi, G.L.; Roshmi, V.; Sameera, Y.V.; Pradeepini, G. Credit Card Fraud Detection using Ensemble Methods in Machine Learning. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1237–1241. [CrossRef]
23. Tian, Y.; Liu, G.; Wang, J.; Zhou, M. ASA-GNN: Adaptive Sampling and Aggregation-Based Graph Neural Network for Transaction Fraud Detection. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 3536–3549. [CrossRef]
24. Xie, Y.; Liu, G.; Yan, C.; Jiang, C.; Zhou, M. Time-Aware Attention-Based Gated Network for Credit Card Fraud Detection by Extracting Transactional Behaviors. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 1004–1016. [CrossRef]
25. Xie, Y.; Liu, G.; Zhou, M.; Wei, L.; Zhu, H.; Zhou, R. A Spatial-Temporal Gated Network for Credit Card Fraud Detection by Learning Transactional Representations. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 6978–6991. [CrossRef]
26. Xie, Y.; Liu, G.; Yan, C.; Jiang, C.; Zhou, M.; Li, M. Learning Transactional Behavioral Representations for Credit Card Fraud Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 5735–5748. [CrossRef] [PubMed]
27. Worldline and the Machine Learning Group of ULB, Credit Card Fraud Detection Dataset. Available online: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (accessed on 10 December 2024).
28. Niu, X.; Wang, L.; Yang, X. A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised. *arXiv* **2019**, arXiv:1904.10604.
29. Uddin, M.F. Addressing accuracy paradox using enhanced weighted performance metric in machine learning. In Proceedings of the ITT 2019-Information Technology Trends: Emerging Technologies Blockchain and IoT, Ras Al Khaimah, United Arab Emirates, 20–21 November 2019; pp. 319–324. [CrossRef]
30. Ndama, O.; Bensassi, I.; En-Naimi, E.M. Optimizing credit card fraud detection: A deep learning approach to imbalanced datasets. *Int. J. Electr. Comput. Eng.* **2024**, *14*, 4802–4814. [CrossRef]
31. Fanai, H.; Abbasimehr, H. A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection. *Expert Syst. Appl.* **2023**, *217*, 119562. [CrossRef]
32. Alsuwailem, A.A.S.; Salem, E.; Saudagar, A.K.J. Performance of Different Machine Learning Algorithms in Detecting Financial Fraud. *Comput. Econ.* **2023**, *62*, 1631–1667. [CrossRef]
33. Feng, X.; Kim, S.K. Novel Machine Learning Based Credit Card Fraud Detection System. *Mathematics* **2024**, *12*, 1869. [CrossRef]
34. Awoyemi, J.O.; Adetunmbi, A.O.; Oluwadare, S.A. Credit card fraud detection using machine learning techniques: A comparative analysis. In Proceedings of the 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 29–31 October 2017; pp. 1–9. [CrossRef]
35. Zhu, H.; Liu, G.; Zhou, M.; Xie, Y.; Abusorrah, A.; Kang, Q. Optimizing Weighted Extreme Learning Machines for imbalanced classification and application to credit card fraud detection. *Neurocomputing* **2020**, *407*, 50–62. [CrossRef]
36. Boyd, K.; Eng, K.H.; Page, C.D. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science*; Blockeel, H., Kersting, K., Nijssen, S., Železný, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8190, pp. 451–466. [CrossRef]
37. Hilal, W.; Gadsden, S.A.; Yawney, J. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Syst. Appl.* **2022**, *193*, 116429. [CrossRef]
38. Tang, Y.; Liang, Y. Credit card fraud detection based on federated graph learning. *Expert Syst. Appl.* **2024**, *256*, 124979. [CrossRef]
39. Zhao, X.; Liu, Y.; Zhao, Q. Improved LightGBM for Extremely Imbalanced Data and Application to Credit Card Fraud Detection. *IEEE Access* **2024**, *12*, 159316–159335. [CrossRef]
40. Asha, R.B.; Suresh Kumar, K.R. Credit card fraud detection using artificial neural network. *Glob. Transit. Proc.* **2021**, *2*, 35–41. [CrossRef]
41. Kalid, S.N.; Ng, K.H.; Tong, G.K.; Khor, K.C. A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes. *IEEE Access* **2020**, *8*, 28210–28221. [CrossRef]
42. Baisholan, N.; Turdalyuly, M.; Gnatyuk, S.; Baisholanova, K.; Kubayev, K. Implementation of Machine Learning Techniques to Detect Fraudulent Credit Card Transactions on a Designed Dataset. *J. Theor. Appl. Inf. Technol.* **2023**, *101*, 5279–5287.