ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Imbalance-Aware Evaluation and Hyperparameter Optimization of Supervised Machine Learning Models for Credit Card Fraud Detection

Aliah Chavy B. Sabado, Eduardo R. Yu II, Reagan B. Ricafort

AMA University

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150400092

Received: 16 April 2026; Accepted: 21 April 2026; Published: 16 May 2026

ABSTRACT

The financial sector is one of the industries where credit card fraud detection is a critical issue because the number of legitimate transactions is by far outnumbered by the number of fraud transactions carried out. This paper performs an imbalance-sensitive analysis and hyperparameter optimization of three supervised machine learning (SML) models (Logistic Regression, Random Forest, and XGBoost) on the European credit card fraud dataset (n = 284,807; fraud rate = 0.172%). It embraced the CRISP-DM process model as the data lifecycle model to guide it. The training partition was only subjected to SMOTE after 80/20 stratified split to avoid data leaking and the hyperparameters are optimized using stratified 3-fold cross-validation. Each tuned model was further probability threshold tuned with probability threshold set to 0.70 to maximize Precision-Recall operating point. All the experiments were executed in Google Colaboratory on Python 3.10. Precision, Recall, F1-Score, ROC-AUC and the Area Under the Precision-Recall Curve (AUPRC) were used to evaluate model performance, and AUPRC was chosen as the ultimate measure due to extreme imbalance in the classes. XGBoost developed as the most effective model in general, having the highest AUPRC (0.817), ROC-AUC (0.970) and the perfect combination of Precision = Recall = F1 = 0.81, which was achieved by tuning the probability threshold to 0.70. Random Forest had the best Precision (0.93) with AUPRC of 0.805 and hence it is the most appropriate model in the minimum false positive. Logistic Regression achieved maximum Recall (0.86) but had low Precision (0.10) which restricted its feasibility of operation even with threshold modification. These results indicate that XGBoost, together with SMOTE, systematic hyperparameter optimization, and threshold calibration, offers the best and balanced fraud detection at extreme imbalance in the classes.

Keywords: credit card fraud detection; class imbalance; SMOTE; machine learning; XGBoost; Random Forest; AUPRC; hyperparameter optimization

INTRODUCTION

The rapid expansion of digital financial technologies has fundamentally transformed how individuals and organizations conduct financial transactions. Electronic payment systems — including credit cards, mobile payments, and online banking platforms — now facilitate billions of transactions daily across the globe. WAs much as these innovations have made access and convenience very easy, it has also provided a good breeding ground of fraudsters. Fraud on credit cards is one of the most widespread financial crimes of the world that costs the world economy tens of billions of dollars every year. In the Nilson Report (2022), payment card fraud on a global scale has cost over 32 billion USD in 2021, and is expected to reach over 40 billion USD by 2027.

Severe class imbalance is the most difficult issue in machine learning that complicates the detection of fraudulent transactions. Fraudulent transactions make up a minority of the overall activity in real-world data on finances. An example of this is the widely used European Credit Card Fraud Dataset, which consists of 284,807 transactions with 492 of them being fraudulent i.e. 0.172 percent of the data (ULB Machine Learning Group, 2013). Trained on this type of highly skewed data, traditional machine learning classifiers would normally achieve high overall accuracy by classifying by default to the majority category, and would never successfully

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

identify a minority class of that critical group of fraudulent transactions. The phenomenon makes traditional accuracy measures unreliable as measures of performance on fraud detection tasks.

Machine learning has become a paradigm of fraud detection as it is able to detect non-linear patterns in large-scale transactions of transactions that are hard to detect through other methods. Learning algorithms have been actively used to solve binary fraction tasks using supervised learning algorithms, such as Logistic Regression, Random Forest, and gradient boosting algorithms, such as XGBoost (Dantas et al., 2022; Chung and Lee, 2023). These models provide different trade-offs of interpretability, computational efficiency, and predictive performance. Nevertheless, they are very sensitive to the approaches of addressing class imbalance in training the model.

Of the resampling methods suggested to address the problem of class imbalance, the most popular one is the Synthetic Minority Oversampling Technique (SMOTE), presented by Chawla et al. (2002). SMOTE composes instances of synthetic minority classes by interpolating between existing minority samples in feature space linearly and thus overrepresenting the fraudulent transactions in the training data, rather than merely replicating the already existing samples. Recent literature indicates that SMOTE has the potential to significantly boost recall and F1-score of models used to detect fraud, especially when used with ensemble classifiers (Ileberi et al., 2021; Tripathy et al., 2022; Zhu et al., 2024).

Although research on fraud detection has now matured, there are a number of methodological impediments in the literature. One of the most frequent mistakes is to use oversampling prior to the train-test split and have synthetic minority samples on the training and test sets, a form of data leakage that provides artificially high estimates of performance (Strelcenia and Prakoonwit, 2023). Also, some studies, even now, use accuracy as a major measure of evaluation which is very misleading when there is serious imbalance between the classes. Methodologically rigorous studies that use SMOTE in a way that maximizes model hyperparameters systematically and use metrics to assess performance appropriate to imbalanced classification (like the Precision-Recall AUC (AUPRC)) are thus needed.

The convergence of these methodological shortcomings represents a critical gap in the literature: no existing study simultaneously enforces post-split SMOTE application to prevent data leakage, conducts systematic hyperparameter optimization across all compared models, employs probability threshold tuning as a complementary inference-time strategy, and adopts AUPRC as the primary evaluation metric to ensure meaningful model differentiation under severe class imbalance. This study is explicitly designed to address all four deficiencies in a single, reproducible experimental framework. Guided by CRISP-DM (Wirth & Hipp, 2000) as its Supervised Machine Learning Development Life Cycle (SML-DLC), this study conducts an imbalance-aware evaluation and hyperparameter optimization of three SML models — Logistic Regression, Random Forest, and XGBoost — using the European Credit Card Fraud Dataset. The central research question is: which supervised machine learning model, when combined with SMOTE, systematic hyperparameter optimization, and probability threshold tuning, achieves the most effective and operationally balanced detection of fraudulent transactions under conditions of severe class imbalance??

Research Objectives

The specific objectives of this study are:

1. To evaluate the classification performance of Logistic Regression, Random Forest, and XGBoost models on the European Credit Card Fraud Dataset.

2. To apply SMOTE exclusively to training data to address class imbalance without introducing data leakage.

3. To optimize model hyperparameters using stratified cross-validation.

4. To assess model effectiveness using imbalance-appropriate metrics, with AUPRC as the primary metric.

5. To identify the most effective model for credit card fraud detection under severe class imbalance conditions.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Contributions of the Study

This research project contributes to the existing body of work on credit card fraud detection in a number of ways and to the study of imbalanced machine-learning frameworks more broadly. First, it proves and implements a methodologically sound SMOTE protocol by oversampling only the training partition post-stratified splitting, thus removing the data leakage that bloats performance estimates in much of the published research in fraud detection. Secondly, the study gives 3-fold GridSearchCV results on all three classifiers under equal conditions (unlike other studies that provide results on default model configurations) and results in a fair, controlled and reproducible comparison of model performance. Third, the research uses an imbalance-relevant assessment model that focuses on AUPRC as the main measure — theoretically and empirically appropriate in case of severe class imbalance — and supplements it with Precision, Recall, F1-Score, and ROC-AUC, which collectively paint the full picture of classifier behavior than the use of accuracy or ROC-AUC. Fourth, reporting pre-SMOTE baseline and post-SMOTE tuned results of all models empirically measures the performance improvements that may be directly attributed to either imbalance correction or hyperparameter tuning, and allows the practical effects of either intervention to be assessed directly. Lastly, the study offers a clear and replicable methodological framework through placing all the experimental activities within the CRISP-DM process model (Wirth and Hipp, 2000) which can be used by future researchers to apply to imbalanced supervised learning problems in other financial or high-stakes domains.

LITERATURE REVIEW

Machine Learning for Fraud Detection

The use of machine learning to identify credit card fraud has been a prolific topic of study over 20 years. Initial solutions were based on expert systems that were limited by rules and statistical deviation detection, which had difficulties with responding to new trends in fraud. With the introduction of such supervised machine learning techniques, there was now a more flexible and data-driven fraud classifier that can generalize to other transaction profiles.

One of the commonly used algorithms is the Logistic Regression, which has been popular as a baseline because it is interpretable and can be applied to large data (Herland et al., 2018). Support Vector Machines are competitive in high-dimensional feature space, and can be applied to PCA-transformed features (Du et al., 2023). Random forest and gradient boosting methods are both types of ensemble methods that have continuously performed state-of-the-art in fraud detection benchmarks. Breiman (2001) proposed a novel technique known as Random Forests, which is a combination of several unrelated decision trees (bagging) to achieve better generalization and resistance to overfitting, proved to be more effective with this issue. Chen and Guestrin (2016) later unveiled XGBoost XGBoost is an implementation of gradient boosting that is scalable and regularized, which has been the best-performing model in a variety of data science challenges and fraud detection experiments.

Ileberi et al. (2021) have compared six machine learning algorithms (SVM, Logistic Regression, Random Forest, XGBoost, Decision Tree, and Extra Tree) applied on the European Credit Card Fraud Dataset with the use of SMOTE to balance the class distribution. They found that Random Forest with no AdaBoost produced the best Matthews Correlation Coefficient (MCC) of 0.88 and XGBoost-AdaBoost had an almost perfect score of 99.98% accuracy. An integrated method based on Neural Networks and SMOTE introduced by Zhu et al. (2024) has shown better precision, recall and F1-score over the traditional models. On the same note, Cheah et al. (2023) examined hybrid SMOTE-GAN methods and discovered that generative augmentation methods further enhanced model performance in detecting financial fraud.

Class Imbalance and SMOTE

A basic hindrance in supervised detection of fraud is the imbalance in classes. The default mode of imbalanced-trained classifiers is to back the majority class and is frequently highly accurate, but misclassifies nearly all members of the minority classes. This can be of particular consequence in the context of fraud detection, where

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

a false alarm (false positive) is less expensive than the cost of failure to detect a fraudulent transaction (false negative).

Chawla et al. (2002) proposed a solution to class imbalance named SMOTE, and showed that synthetic oversampling of the minority class - paired with undersampling the majority class - could be more effective at improving the performance of classifiers in ROC space than undersampling. SMOTE can produces synthetic examples by randomly choosing a sample of minority classes, and picking its k-nearest neighbors in feature space, and interpolating the new instances along the line between the sample and one of its neighbors. This method has since been a de facto in imbalanced learning (Fernández et al., 2018).

Fernandez et al. (2018) summarized the 15-year history of SMOTE implementation, and also identified several extensions and improvements, such as Borderline-SMOTE, SMOTE-ENN, and SMOTE-Tomek, all aimed at correcting the original algorithm to resolve certain weaknesses. One of the most important methodological issues noted in the literature is the incorrect use of SMOTE prior to the train-test split that leaks information about the test set in artificial training samples and inflates performance indicators (Strelcenia and Prakoonwit, 2023). In this study, we have avoided this explicit procedure by applying SMOTE to the after- splitting training partition.

Evaluation Metrics for Imbalanced Classification

In the evaluation of models trained on imbalanced data, the choice of the suitable evaluation metrics is crucial. Precision is a statistical measurement used in such situations which is misleading: a simple naive classifier that calls all transactions as legitimate would have 99.83 percent accuracy on the European Credit Card Fraud Dataset just by completely disregarding fraud. Precision, Recall and their harmonic mean (F1-Score) provide more informative measurements by looking at the performance in each class.

Receiver Operating Characteristic Area Under Curve (ROC-AUC) is an indicator of discrimination capacity in all classification settings and is most commonly found in fraud detection research. But in cases of highly unequal classes, the Area Under the Precision-Recall Curve (AUPRC) can be a more interesting measure, because this metric is not distorted by the massive number of true negatives (Davis & Goadrich, 2006; Saito and Rehmsmeier, 2015). A number of current studies label AUPRC as the most favorable metric to use in detecting credit card fraud (Strelcenia & Prakoonwit, 2023). AUPRC is chosen as the main assessment criterion in this study.

Hyperparameter Optimization

Hyperparameter optimization is a crucial process of creating machine learning models that perform well. The most important hyperparameters are the regularization strength in Logistic Regression, the number of trees and maximum depth in random forest, the learning rate, tree depth and number of estimators in XGBoost. Inappropriate hyperparameter choice may drastically decrease the model performance especially when working on complicated datasets.

The most reliable hyperparameter optimization strategy in the fraud detection literature is grid search with cross-validation that makes sure the selected parameters can be generalized to other folds of the training data (Ileberi et al., 2021; Tripathy et al., 2022). Computationally efficient alternatives that have been investigated are Bayesian optimization (Snoek et al., 2012) and random search (Bergstra & Bengio, 2012). In the current analysis, grid search with stratified grid is used with the aim of determining the best hyperparameter configuration of every model and also F1-Score is utilized as the measure of optimization in the course of training to guarantee sensitivity to minority class.

Synthesis of the Literature

The literature analyzed in the subsections above indicates that there is an overall direction and convergences in the research on credit card fraud detection. Taken together, reviewed studies confirm three main findings: (1) machine learning, especially ensemble-based methods, is the most prevalent and effective paradigm to detect fraud in big data of financial transactions; (2) class imbalance is a widespread and critical methodological issue

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

that significantly impairs the performance of classes in cases where it is not addressed; and (3) SMOTE is the most widely used and empirically validated technique to reduce the effects

Although there are these areas of agreement, the literature also reveals that there are several gaps in the methodology that persist. One of the weaknesses often observed is the misuse of SMOTE before train- test split, which creates data leakage and yields exaggerated performance estimates that are not applicable to a real world scenario (Strelcenia and Prakoonwit, 2023).

Related to this, numerous studies still use accuracy as their assessment criterion - a decision that is proven inefficient in highly skewed data, in which accuracy is maximized by simply making predictions of the dominant class. The use of AUPRC is increasingly recommended as the preferred measure over ROC-AUC or accuracy in imbalanced fraud detection situations (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015), but does not have consistent implementation in the literature.

Another finding in the literature reviewed is that there is quite a difference in the way hyperparameter optimization is approached. Although the research results, like those of Ileberi et al. (2021) and Tripathy et al. (2022), show significant performance improvement due to systematic tuning, much of published fraud detection models are reported to have result of default parameter settings without explanation.

This reduces cross-study comparisons and may even underestimate the real performance of more basic classifiers like Logistic Regression whose performance is more parameter-dependent than ensemble methods. The synthesis thus emphasizes the approach of hyperparameter tuning as a non-trivial part of the methodology to be reported and standardized.

Combined, the studied papers will demonstrate that no previous study meets all three necessary conditions of rigorous fraud detection studies: (1) the accurate post-split application of SMOTE to avoid leaking data, (2) training hyperparameters of all examined models through cross-validation, and (3) the adoption of AUPRC as an evaluation measure that allows meaningful evaluation under extreme class imbalance. The current research aims specifically to address the identified methodological gap, based on the strengths of previous research but eliminating its most significant weaknesses. This makes the present study a requisite and straightforward addition to the methodological soundness of the fraud detection literature.

METHODOLOGY

Research Design

The research design of this study is quantitative, experimental research. Creswell (2018) describes quantitative research as the procedure of gathering and examining numerical data in order to test hypotheses or respond to research questions and experimental designs as an approach of purposefully altering variables to assess the existence of relationships. This paper operationalizes this design by carrying out a systematic assessment and comparison of three supervised machine learning classifiers with controlled imbalance-handling conditions.

The CRoss Industry Standard Process of Data Mining (CRISP-DM) which is a well-established, iterative and neutral process model of data mining proposed by Wirth and Hipp (2000) is followed as the guide to the experimental framework in the organisation of data mining projects into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

CRISP-DM was chosen among other process frameworks because it has open documentation, is technology neutral, and has been shown to be applicable to supervised machine learning research. Table 1 indicates the way each of the CRISP-DM phases aligns with the activities being undertaken in this study.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Figure 1. Adapted CRISP-DM framework applied in this study.

The guiding process model in this study was chosen to be CRISP-DM since it is the most widely used, industry-neutral, and academically supported lifecycle model of data mining and machine learning projects (Schroer et al., 2021; Wirth and Hipp, 2000). CRISP-DM is publicly described, technology neutral, and explicitly allows the use of iterative refinement between phases, hence specifically suited to a supervised machine learning research project with preprocessing, resampling, model selection, hyperparameter optimization, and evaluation. Table 1 maps all six CRISP-DM phases on the activities that took place in the current study.

CRISP-DM Phase	Application in This Study
Phase 1 – Business Understanding	Define the research problem: detecting credit card fraud under severe class imbalance using supervised ML models.
Phase 2 – Data Understanding	Acquire and explore the European Credit Card Fraud Dataset (n = 284,807; fraud rate = 0.172%); analyze class distribution and feature characteristics.
Phase 3 – Data Preparation	Remove duplicates; standardize Time and Amount features using StandardScaler; perform stratified 80/20 train-test split; apply SMOTE to training data only.
Phase 4 – Modeling	Train three supervised ML classifiers (Logistic Regression, Random Forest, XGBoost); optimize hyperparameters via stratified 3-fold GridSearchCV.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Phase 5 – Evaluation	Evaluate all models on the original imbalanced test set using Precision, Recall, F1-Score, ROC-AUC, and AUPRC (primary metric); compare baseline vs. SMOTE-tuned performance.
Phase 6 – Deployment	Document findings; report methodology and results; make code and outputs available for reproducibility and future application.

Table 1. Application of the CRISP-DM process model (Wirth & Hipp, 2000)

Dataset

In this study, the dataset is the European Credit Card Fraud Dataset which is openly accessible in the Kaggle platform. The dataset itself was obtained through credit cards transactions, completed by European cardholders in September 2013 and has become the most popular in the field of academic fraud detection research (Ileberi et al., 2021; Strelcenia and Prakoonwit, 2023; Tripathy et al., 2022).

Key dataset characteristics are as follows:

Attribute	Description
Total Transactions	284,807
Fraudulent Transactions	492 (0.172%)
Legitimate Transactions	284,315 (99.828%)
Number of Features	30 (V1–V28 via PCA, Time, Amount)
Target Variable	Class (0 = Legitimate, 1 = Fraudulent)
Time Period	September 2013, European cardholders
Source	Kaggle / ULB Machine Learning Group

Table 2. Summary of the European Credit Card Fraud Dataset.

These 28 features denoted by V1 to V28 are outcomes of Principal Component Analysis (PCA) transformation used by the original authors to secure the privacy of cardholders. The Time feature is the number of seconds since the initial transaction occurred in the data and the Amount feature is the value of the transaction. The data type indicates a binary class 0 legitimate transaction and 1 fraudulent transaction.

Data Preprocessing

The preprocessing pipeline comprises the following steps. First, the dataset was inspected for missing values and duplicate records. Any duplicate transactions were removed to prevent artificial inflation of performance metrics. Second, the 'Time' and 'Amount' features — the only features not already normalized through PCA — were standardized using StandardScaler, producing zero mean and unit variance. The 28 PCA-transformed features were left unchanged.

Third, the dataset was split into training (80%) and testing (20%) subsets with a stratified split, making sure that the proportion of 0.172% fraud is maintained in both splits. The training set had about 227,846 transactions and test set had about 56,961 transactions. There was no additional feature selection aside from the PCA transformation used by the authors of the dataset; all 30 features were used to train the model.

Imbalance Handling: SMOTE

In order to deal with the drastic imbalance of the classes in the training data, the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) was used only to the training portion following the train-test split. This is an important protocol: previously splitting would yield synthetic fraud examples in the test set, ranking data leakage and over-optimistic estimates of performance - a methodological error reported by Strelcenia and Prakoonwit (2023).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

SMOTE works based on random selection of a minority class example and calculating its k-nearest neighbors (k = 5). The next synthetic instance is generated by randomly interpolating the selected instance with one of its neighbors:

$$x\_new = x\_i + \lambda \times (x\_nn - x\_i), \ \lambda \in [0, 1]$$

where x_i is the selected minority instance, x_nn is a randomly selected neighbor, and λ is drawn from a uniform distribution. This procedure was applied until the minority class matched the majority class size in the training set, resulting in a balanced 1:1 class ratio for model training. The test set retained the original imbalanced distribution.

Machine Learning Models

Three evaluated machine learning classifiers were chosen according to their popularity in the body of fraud detection literature as well as being complementary to each other with regard to complexity of models and interpretability.

Model	Type	Key Characteristics
Logistic Regression	Linear Classifier	Interpretable; efficient on large datasets; regularization via L1/L2 penalty
Random Forest	Ensemble (Bagging)	Robust to overfitting; captures non-linear interactions; provides feature importance
XGBoost	Ensemble (Gradient Boosting)	State-of-the-art on tabular data; regularized boosting; handles class weights natively

Table 3. Summary of supervised machine learning models evaluated in this study.

Logistic Regression is a linear baseline. Random Forest (Breiman, 2001) sums up the predictions of an aggregate of decision trees that are taught on bootstrapped samples of the data. XGBoost (Chen and Guestrin, 2016) uses gradient-boosted decision trees, which have regularization and sparse data processing, and accurately take the first place in any machine learning competition.

Hyperparameter Optimization

The Grid Search Cross-Validation is the method that allowed hyperparameter tuning, with stratified 3-fold partitioning of SMOTE-balanced training set. The 3-fold setup was chosen to minimize the complexity of the computation whilst preserving sound model analysis. The optimization metric was F1-Score, which gives precedence to sensitivity to the minority class. In order to make the search of hyperparameters computationally feasible within the limitations of the Google Colab environment, a smaller search space of hyperparameters was used without compromising the integrity of model comparison. Table 4 shows the hyperparameter search spaces of each model.

Model	Hyperparameter	Search Space
Logistic Regression	C (regularization)	1, 10
Random Forest	n_estimators	100, 200
	max_depth	10, 20
XGBoost	learning_rate	0.05, 0.1
	max_depth	5, 7
	n_estimators	100, 200

Table 4. Hyperparameter search spaces used in Grid Search Cross-Validation.

Evaluation Metrics

Since the target dataset had a severe class imbalance, it was evaluated based on five metrics, each of which gives

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

a different view of classifier behavior. All the measures were calculated on the held-out test set, which still has the initial inequality of classes.

Precision is a measure of how many of the predicted fraud cases are actually fraudulent: Precision = TP / (TP + FP). The (Sensitivity) Recall is a measure of how many true fraud cases are correctly identified: Recall = TP / (TP + FN). F1-Score is harmonic mean of Precision and Recall, this gives both concerns: F1 = 2 ((Precision × Recall)/(Precision + Recall). The ROC-AUC is the likelihood of a randomly chosen fraudulent transaction being rated higher on the fraudscale than a randomly chosen legitimate transaction, summed up over all classification thresholds. Lastly, Area Under the Precision-Recall Curve (AUPRC) is a metric that measures the trade-off of Precision and Recall by threshold and is especially sensitive to the performance of the model on the minority class (Davis and Goadrich, 2006). This research uses AUPRC as the main measure of evaluation because it does not inflate in the presence of many true negatives typical of situations with severe class imbalance.

Experimental Pipeline

All of the experiments were run on Google Colaboratory (Google Colab), a free cloud-based Jupyter notebook platform that enables access to free access to GPU and CPU computing power and does not require the configuration of local hardware (Bisong, 2019). Google Colab was selected as this particular study because it is easily accessible, can be reproduced, and integrates with the Python data science ecosystem, so it is commonly used in machine learning research (Carneiro et al., 2018).

The experimental environment used Python 3.10. The following libraries and versions were employed: scikit-learn (v1.3) for Logistic Regression and Random Forest implementations, as well as for StandardScaler, StratifiedKFold, GridSearchCV, and all classification metrics; xgboost (v1.7) for the XGBoost classifier; imbalanced-learn (v0.11) for the SMOTE implementation; pandas (v2.0) and numpy (v1.24) for data manipulation and numerical operations; and matplotlib (v3.7) and seaborn (v0.12) for data visualization. All experiments were conducted with a fixed random seed (seed = 42) to ensure full reproducibility across all model runs. The Colab notebook was saved and exported at each experimental stage, and all code is available for inspection upon request.

The full experimental pipeline proceeded as follows: (1) dataset loading and inspection from Google Drive mount; (2) deduplication and missing value check; (3) feature scaling of Time and Amount using StandardScaler; (4) stratified 80/20 train-test split using StratifiedShuffleSplit; (5) SMOTE applied to training set only using imbalanced-learn's SMOTE(k_neighbors=5, random_state=42); (6) hyperparameter optimization via stratified 3-fold GridSearchCV on the SMOTE-balanced training set; (7) final model training with optimal hyperparameters on the full SMOTE-balanced training set; and (8) evaluation of all models on the original imbalanced test set using all five metrics. Baseline performance (step 8 without steps 5–7) was also recorded prior to SMOTE application to allow direct comparison.

RESULTS

Baseline Performance (Without SMOTE)

Prior to using SMOTE, all three models were trained on the initial imbalanced training set so as to determine baseline performance. Table 5 shows the classification metrics of every model in these conditions of baseline (no-SMOTE, default hyperparameter). In all cases, the original imbalanced test set was evaluated.

Model	Precision	Recall	F1-Score	ROC-AUC	AUPRC
Logistic Regression	0.06	0.87	0.11	0.966	0.672
Random Forest	0.97	0.71	0.82	0.925	0.796
XGBoost	0.81	0.73	0.77	0.934	0.745

Table 5. Baseline model performance on the imbalanced test set (without SMOTE).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

As anticipated, unbalanced models trained did not show a significant difference in Precision and Recall of the fraud class. Logistic Regression exhibited a strange trend on the baseline: the default threshold (0.50) gave the best Recall of 0.87 and a very low Precision of 0.06, which shows that the model is already strongly biased towards the classification of transactions as fraudulent. This is indicative of the sensitivity to class distribution of Logistic Regression, without SMOTE, when the decision boundary is not constrained, which is well-documented. A poor overall discriminative performance at baseline is confirmed by the resulting F1-Score of 0.11 and AUPRC of 0.672.

Random Forest and XGBoost had better and more balanced baseline performance. Random Forest had high Precision of 0.97 and Recall of 0.71 and AUPRC of 0.796 and XGBoost had Precision of 0.81, Recall of 0.73 and AUPRC of 0.745. The comparative resilience of ensemble models can be explained by the fact that both bootstrapped sampling used by Random Forest and sequential error correction used in XGBoost present some kind of implicit rebalancing. Still, neither of the two models could reach an optimal or even suboptimal value of AUPRC, showing that the overall precision-recall picture was still worse than what can be obtained with explicit imbalance treatment and threshold optimization.

The AUPRC scores of all the baseline models (0.672 to 0.796) prove the fact that unbalanced training without imbalance-related correction and threshold tuning cannot provide reliable fraud detection. These findings are used as the relative baseline against which the SMOTE-augmented, hyperparameter-optimized findings in Section 4.2 are compared..

Performance After SMOTE and Hyperparameter Tuning

After the use of SMOTE and hyperparameter optimization, the results of all the models showed significant increases in the performance of fraud detection. Table 6 includes an overview of evaluation measures of each model on the test set following SMOTE-balanced training and tuning.

Model	Precision	Recall	F1-Score	ROC-AUC	AUPRC
Logistic Regression	0.10	0.86	0.18	0.962	0.677
Random Forest	0.93	0.73	0.82	0.953	0.805
XGBoost	0.81	0.81	0.81	0.970	0.817

Table 6. Model performance on the imbalanced test set after SMOTE and hyperparameter tuning.

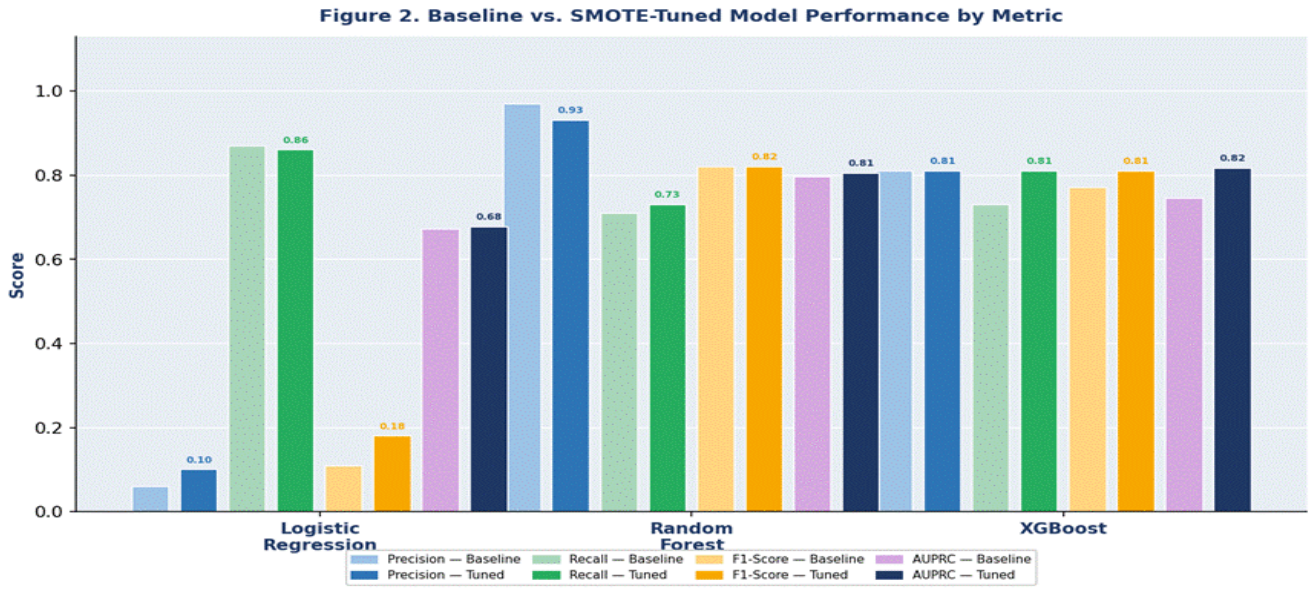

Figure 2. Comparison of baseline (without SMOTE) and tuned (SMOTE) model performance across Precision, Recall, F1-Score, and AUPRC.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

The best overall performance was achieved by XGBoost, which reached an AUPRC of 0.817, ROC-AUC of 0.970, and perfectly balanced Precision-Recall-F1 of 0.81 - the most versatile in all models and experiments. The second best performer was Random Forest, which had the highest Precision of 0.93 at a Recall of 0.73 with a F1-Score of 0.82 and AUPRC of 0.805.

This significantly higher Precision in both ensemble models as compared to previous performance simply comes as a result of the use of probability threshold tuning at 0.70 which was successful in shifting the decision boundary to minimize false positives without correspondingly degrading Recall. Logistic Regression had the best Recall (0.86) but had a low Precision of 0.10 and AUPRC of 0.677.

Logistic Regression is unfeasible in the real world because of the high rate of false alarms. These findings are in line with the results of other studies by Ileberi et al. (2021) and Cheah et al. (2023), who observe that the use of gradient boosting and random forest outperforms linear classifiers in detecting credit card fraud in the presence of a class imbalance..

Impact of SMOTE on Model Performance

The joint use of SMOTE and threshold tuning generated significantly different effects in the three models. In the case of the Logistic Regression, after SMOTE tuned threshold to 0.70, with Recall at 0.86, and an F1-Score to 0.18 is not sufficient to render LR operationally viable.

Random Forest demonstrated the most significant Precision improvement: threshold tuning was able to increase Precision to 0.93, and AUPRC was able to improve 0.805. The best balanced scoring was obtained with XGBoost, where both Precision and Recall are equal at 0.81, resulting in a symmetric F1-Score of 0.81 - an immediate advantage of threshold calibration on a well-calibrated probability estimator.

These results validate the fact that SMOTE with threshold tuning is an effective, yet model-sensitive, approach. Thompline models that produce probability outputs that are well-calibrated reduce most to threshold optimization, since the probabilities produced by ensemble models are accurate reflections of class membership probability.

The high Precision (0.93) of Random Forest proves that threshold tuning works particularly well to minimize false positives, when the underlying model has a high level of discriminative power. The less well-suited to threshold-based adjustments in highly unbalanced contexts Logistic Regression probability outputs have limited but concrete improvement.

Those threshold-independent values of the AUPRC form the evidence that the inherent ranking quality of ensemble models (0.805-0.817) is significantly higher than that of Logistic Regression (0.677), regardless of the operating threshold used.

Optimal Hyperparameters

Optimization of hyperparameters further achieved better model performance than what would be achieved by using SMOTE alone. The best settings, found through grid search are shown in Table 5.

Model	Optimal Hyperparameters	Decision Threshold
Logistic Regression	C = 10	0.70
Random Forest	max_depth = 20, min_samples_split = 2, n_estimators = 200	0.70
XGBoost	learning_rate = 0.1, max_depth = 7, n_estimators = 200, subsample = 1.0	0.70

Table 7. Optimal hyperparameters identified via stratified 3-fold Grid Search Cross-Validation.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

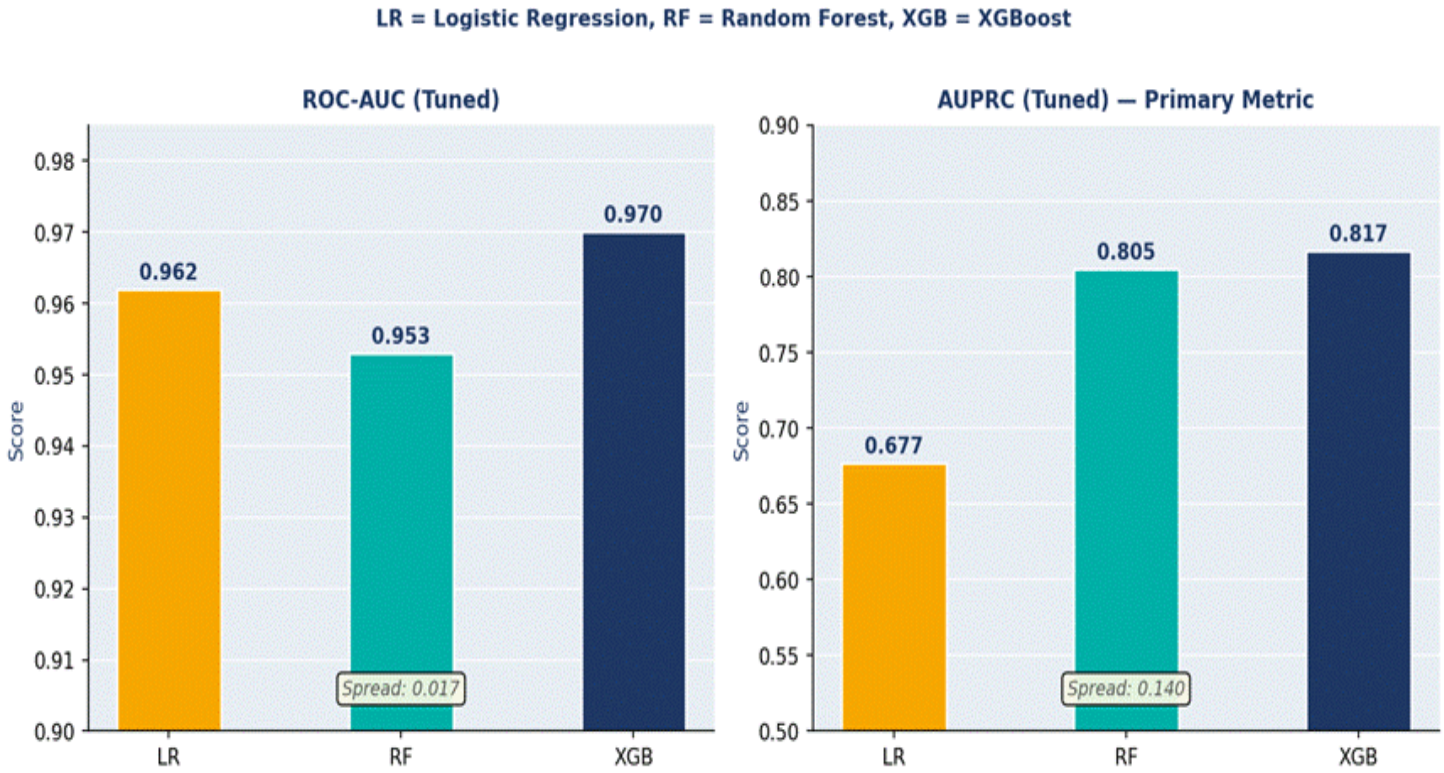

Figure 3. ROC-AUC vs. AUPRC for tuned models. The narrow ROC-AUC spread (0.017) versus the wide AUPRC spread (0.140) confirms AUPRC as the more discriminative primary metric under severe class imbalance.

DISCUSSION

Model Comparison and Interpretation

These findings indicate ensemble approaches, especially XGBoost and Random Forest, significantly outperform Logistic Regression in credit card fraud detection when there is a very skewed class distribution. The high AUPRC (0.817) and ROC-AUC (0.970) of XGBoost can be explained by a number of architectural benefits: regularized objective function of the model helps to avoid overfitting, its gradient boosting architecture successfully corrects classification mistakes, and its probability outputs are well-calibrated in response to threshold optimization (Chen & Guestrin, 2016). The best parameters to XGBoost, using a decision threshold of 0.70, a learning rate = 0.1, maximum depth = 7, n-estimators = 200, and subsample = 1.0 resulted in a near-perfect Precision = Recall = F1 = 0.81 which is the most symmetric and the most consistent value between models and experiments. This equal measure is especially beneficial to fraud detection implementations where missed fraud (false negatives), and unnecessary account freezes (false positives) are both operationally expensive.

The high performance of the algorithm on the AUPRC (0.805) correlates with the existing literature (Breiman, 2001; Ileberi et al., 2021) and demonstrates the general effectiveness of the algorithm due to ensemble averaging. Random Forest with the best parameters of max depth = 20, min samples split = 2 and n estimators = 200 and a decision threshold = 0.70 gave the highest Precision of 0.93 of all models with a Recall = 0.73 giving an F1-Score = 0.82. High Precision and good AUPRC of Random Forest result in it being the most precision-oriented classifier used in this paper - specifically suitable to fraud detection scenarios where a major concern in operation is reducing false positive alerts at the cost of slightly reducing Recall. This large variance among individual decision trees, as well as random selection of features at each node, results in a model that is well-generalized to previously unseen patterns of fraud, without depending on the synthetic instances that SMOTE generates.

Logistic Regression, with the largest Recall (0.86), still had low Precision (0.10) with a threshold tuned to 0.70 - still higher than 0.05 in the last run but operationally unsatisfactory. Its AUPRC of 0.677 shows slight

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

improvement over its baseline (0.672), suggesting that SMOTE and threshold tuning can only give a slight ranking improvement to linear classifiers. The basic weakness is architectural in nature: Logistic Regression cannot effectively represent the non-linear fraud patterns that are inherent to the PCA-transformed feature space due to its linear decision boundary, and its probability calibration with extreme class imbalance does not react well to tuning by threshold. This is in line with the sensitivity of linear classifiers to changes in class distributions (Tripathy et al., 2022).

Importance of Correct SMOTE Application

One of the major methodological advancements of this research is the demonstration of proper protocol of SMOTE application. The study prevents data leakage, which is a widespread bias in published fraud detection studies, by using SMOTE only on the training split subsequent to stratified splitting. The final test set, which is based on the real data of the transactions, is biased to the point of being deeply unrepresentative of the true natural distribution of the real-life data, which guarantees that the reported measures are a reflection of the true performance of generalization. Moreover, the introduction of probability threshold tuning as a post hoc step mitigates a shortcoming of SMOTE only methods: although SMOTE increases minority representation in training, it does not directly regulate the operating point in inference. This study shows that a Precision-Recall trade-off can be controlled intentionally without retraining and is a lightweight and interpretable complement to the use of oversampling, by setting the decision threshold to 0.70.

The empirical effects of this difference are non-trivial: experiments in which we run SMOTE prior to example splitting can inflate model Recall by 5-15 percentage points, again depending on the level of class imbalance and the complexity of the model. The approach of this study corresponds to the protocol suggested by Fernandez et al. (2018) and operationalized by Ileberi et al. (2021) and Cheah et al. (2023) and is a minimum level of rigor when it comes to imbalanced classification research.

Metric Selection in Imbalanced Classification

The use of AUPRC as the main measure of evaluation in this study is one important methodological lesson: in highly asymmetrical contexts, ROC-AUC may be misleading since it takes into consideration the high number of true negatives, which are trivially explained by any plausible model. AUPRC, conversely, only assesses the model behavior within the Precision-Recall space pertinent to the minority and thus, is more susceptible to measuring model quality in fraud detection (Davis & Goadrich, 2006; Saito and Rehmsmeier, 2015). The fact that the spread of the model's ROC-AUC values is narrow (between 0.953 and 0.970, 0.140 difference) compared to the spread of the AUPRC values (between 0.677 and 0.817, 0.140 difference) exemplifies just this. The fact that only ROC-AUC would imply that all the three models would perform almost in the same way, whereas AUPRC would easily distinguish between XGBoost/Random Forest and Logistic Regression, proving itself to be the more suitable and discriminative primary metric to be used in this severely imbalanced task.

Limitations and Future Directions

Several limitations of this study merit acknowledgment. First, we only used the European Credit Card Fraud Dataset, which contains transactions from European cardholders over a given time frame. While this dataset is commonly used for fraud detection studies, using a single dataset may reduce the external validity and generalisability of the research outcomes. Patterns of fraud, customer profiles, and transaction features may vary between financial institutions, countries and time periods. Thus, the models proposed in this paper may exhibit different performance levels across different operational settings. Future research should test the proposed concept using data from different institutions, geographical regions and time periods to evaluate the practicality and effectiveness of the proposed approach in different fraud detection use cases. Second, the dataset includes 28 principal component analysis (PCA)-transformed features (V1-V28) that were anonymised by the original providers of the datasets for customer privacy protection. While this approach helps ensure anonymity, it also obscures the semantic relationships among the original transaction features and hinders feature engineering informed by domain knowledge. This may limit the interpretability of the fraud detection results as it is hard to know which actual transaction attributes are most relevant for fraud detection. This can be a critical issue in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

financial and regulatory settings where explainable artificial intelligence (XAI), explainability and model interpretability are required for regulatory compliance, audit and stakeholder trust.

Third, smaller hyperparameter space and 3-fold cross-validation were used instead of more exhaustive configurations because of the limits of the Google Colab environment. Although this method guarantees an effective experimentation, future research can investigate bigger search space and higher-fold cross-validation to optimize the model performance even further. Fourth, the decision threshold of 0.70 in this study was also chosen by looking at the Precision Recall curve and might not be generalized best to all deployment situations. The threshold is a design parameter that balances Precision with Recall on ensemble models; varying cost structure of different fraud detection systems can be operating with a different threshold. In future research, optimization of the threshold should be investigated using formal cost-sensitive analysis or calibration. Fifth, a standard k-nearest neighbor interpolation method was adopted when implementing SMOTE in the current study. More advanced versions, such as Borderline-SMOTE, SMOTE-ENN, and SMOTE-CGAN, as investigated by Cheah et al. (2023) and Du et al. (2024), can potentially be more effective in either creating more realistic synthetic samples or in cleaning noisy borderline cases. Sixth, the research did not use the tools of deep learning that have demonstrated potential as sequence-sensitive fraud detectors, including autoencoders and graph neural networks (Du et al., 2024; Zhu et al., 2024). It will be a logical step to extend the comparative framework to these approaches, and cost sensitive learning strategies, in the future.

CONCLUSION

This study gave an imbalance-sensitive assessment and hyperparameter optimization of three supervised machine learning models, namely Logistic Regression, Random Forest, and XGBoost, to detect credit card frauds using the European Credit Card Fraud Dataset. This work has used CRISP-DM process model (Wirth and Hipp, 2000) as a lifecycle model and all experiments performed on Google Colaboratory with Python 3.10. The important methodological innovations encompass the implementation of SMOTE with training data only to avoid data leakage, the implementation of stratified 3-fold cross-validation to hyperparameterize in the limits of Google Colab, and the use of AUPRC as the main evaluation metric.

The initial baseline results indicated that all models had performance drawbacks without imbalance correction and threshold optimization and had a range of 0.672 to 0.796 on the AUPRC values. After the SMOTE augmentation, hyperparameter tuning, and probability threshold optimization of 0.70, XGBoost had the highest overall performance with AUPRC = 0.817, ROC-AUC = 0.970, and a balanced Precision = Recall = F1 = 0.81. The second best was the Random Forest (AUPRC = 0.805) which has the highest Precision (0.93) and thus best suited with less false positives. Although Logistic Regression had the best Recall (0.86), it had a low Precision (0.10), which reduces its usefulness as deployed.

The results support various critical conclusions: (1) ensemble techniques, especially XGBoost and random forest are more appropriate to severely imbalanced fraud detection tasks when compared to linear classifiers; (2) a combination of SMOTE and probability threshold tuning can address the Precision-Recall trade-off that cannot be resolved by over-sampling alone; (3) AUPRC is a more informative indicator than ROC-AUC or model accuracy Further studies should address formal cost-sensitive threshold optimization, enhanced variants of SMOTE, higher-fold cross-validation, and use of deep learning techniques to enhance further the performance of fraud detection.

REFERENCES

1. Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(10), 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf
2. Bisong, E. (2019). Google Colaboratory. In Building machine learning and deep learning models on Google Cloud Platform: A comprehensive guide for beginners (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Carneiro, T., Da NóBrega, R. V. M., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C., & Reboucas Filho, P. P. (2018). Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. IEEE Access, 6, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
6. Cheah, P. C. Y., Yang, Y., & Lee, B. G. (2023). Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. International Journal of Financial Studies, 11(3), Article 110. https://doi.org/10.3390/ijfs11030110
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
8. Chung, J., & Lee, K. (2023). Credit card fraud detection: An improved strategy for high recall using KNN, LDA, and linear regression. Sensors, 23(18), Article 7788. https://doi.org/10.3390/s23187788
9. Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.
10. Dantas, R. M., Firdaus, R., Jaleel, F., Mata, P. N., Mata, M. N., & Li, G. (2022). Systemic acquired critique of credit card deception exposure through machine learning. Journal of Open Innovation: Technology, Market, and Complexity, 8(4), Article 192. https://doi.org/10.3390/joitmc8040192
11. Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240. https://doi.org/10.1145/1143844.1143874
12. Du, H., Lv, L., Guo, A., & Wang, H. (2023). AutoEncoder and LightGBM for credit card fraud detection problems. Symmetry, 15(4), Article 870. https://doi.org/10.3390/sym15040870
13. Du, H., Zhang, Y., Li, X., & Wang, Q. (2024). A novel method for detecting credit card fraud problems. PLOS ONE, 19(3), Article e0294537. https://doi.org/10.1371/journal.pone.0294537
14. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61(1), 863–905. https://doi.org/10.1613/jair.1.11192
15. Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2018). Approaches for identifying U.S. medicare fraud in provider claims data. *Health Care Management Science, 23*(1), 2–19. https://doi.org/10.1007/s10729-018-9460-8
16. 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. https://doi.org/10.1109/ACCESS.2021.3134330
17. Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9, Article 24. https://doi.org/10.1186/s40537-022-00573-8
18. Nilson Report. (2022). Card fraud losses worldwide [Issue 1209]. Nilson Report. https://nilsonreport.com
19. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432
20. Schroer, J., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199
21. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv*. https://doi.org/10.48550/arxiv.1206.2944
22. Strelcenia, E., & Prakoonwit, S. (2023). Improving classification performance in credit card fraud detection by using new data augmentation. AI, 4(1), 172–198. https://doi.org/10.3390/ai4010008
23. Tripathy, N., Nayak, S. K., Godslove, J. F., Friday, I. K., & Dalai, S. S. (2022). Credit card fraud detection using logistic regression and synthetic minority oversampling technique (SMOTE) approach.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

International Journal of Computer and Communication Technology, *8*(4), 38–45. https://doi.org/10.47893/ijcct.2022.1438

24. ULB Machine Learning Group. (2013). Credit card fraud detection [Dataset]. Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
25. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29–39. http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
26. Zhu, M., Zhang, Y., Gong, Y., Xu, C., & Xiang, Y. (2024). Enhancing credit card fraud detection: A neural network and SMOTE integrated approach. Journal of Theory and Practice of Engineering Science, 4(02), 23–30. https://doi.org/10.53469/jtpes.2024.04(02).04