Fraud Identification in Online Financial Transactions Combining SMOTE and Ensemble Classifiers

Juvinal Ximenes Guterres
Department of Electrical Engineering, State University of Malang, Malang, Indonesia | Oriental University of Timor-Leste, Timor-Leste
juvinal.ximenes.2305349@students.um.ac.id

Triyanna Widwaningtyas
Department of Electrical Engineering, State University of Malang, Malang, Indonesia
triyannaw.ft@um.ac.id (corresponding author)

Ilham Ari Elbaith Zaeni
Department of Electrical Engineering, State University of Malang, Malang, Indonesia
ilham.ari.ft@um.ac.id

Received: 30 November 2025 | Revised: 16 January 2026 | Accepted: 27 January 2026

 DOI: https://doi.org/10.48084/etasr.16615

ABSTRACT

The increasing reliance on digital payment systems has intensified the exposure of financial institutions to credit card fraud, particularly under conditions of extreme class imbalance where fraudulent transactions account for less than 0.2% of total records. This study presents SMOVO, a Synthetic Minority Oversampling Technique (SMOTE)-based Voting Ensemble. It is a fraud detection framework designed to address this imbalance through the integration of three complementary components: Analysis of Variance (ANOVA) F-test-based feature selection, SMOTE-based data resampling, and a heterogeneous soft-voting ensemble comprising Random Forest (RF), XGBoost, LightGBM, and AdaBoost. Evaluated on the European Credit Card Fraud Dataset, SMOVO attained a recall of 0.962, F1-score of 0.864, Area Under the Curve (AUC) of 0.982, Matthews Correlation Coefficient (MCC) of 0.868, and Geometric mean (G-mean) of 0.980. In comparison with generative oversampling approaches such as ESMOTE-GAN, SMOVO exhibits more stable training behavior, reduced computational overhead, and improved model interpretability. These results indicate that SMOVO provides an effective and practically deployable solution for fraud detection in highly imbalanced financial transaction environments.

Keywords-fraud detection; SMOTE; ensemble learning; voting classifier; imbalanced data; machine learning

I. INTRODUCTION

Digital transformation in finance has propelled credit cards to the forefront of non-cash payments worldwide. This digital transformation not only provides more practical, faster, and efficient transaction processes but has also contributed to a substantial increase in cross-border transaction volumes [1]. Nevertheless, behind this positive momentum lies an increasingly complex set of security challenges, particularly the evolving forms of digital financial transaction fraud. Recent data from The Nilson Report (2023) revealed that global losses associated with payment card misuse have surpassed USD 32 billion, with a consistently rising annual trend [2, 3]. This phenomenon is further intensified by the rapid adoption of digital banking, electronic wallets, and application-based payment platforms; whereas these services expand financial access, they concurrently enlarge the attack surface and increase exposure to increasingly sophisticated cyber threats [4]. Findings from the Association of Certified Fraud Examiners (ACFE) also confirm a consistent increase in both the volume and the economic impact of digital financial crimes [5].

Contemporary fraud schemes frequently replicate legitimate transaction patterns, enabling them to evade conventional rule-based detection systems [6]. Criminal economic motives, including illicit transactions that exploit victims' financial identities, necessitate more robust detection models capable of withstanding continually evolving attack patterns [7]. Credit card fraud, including identity theft, card-present fraud, card-not-present fraud, and the compromise of authentication credentials, has become an increasingly complex phenomenon

that can only be effectively addressed through intelligent Machine Learning (ML)-based analytical approaches [8].

ML offers a strategic solution, capable of extracting latent patterns from large-scale data often overlooked by conventional methods, effectively addressing the evolving dynamics of digital transactions [9]. Approaches include supervised learning (e.g., tree-based algorithms and boosting), unsupervised learning (e.g., autoencoders and Isolation Forest), and ensemble or hybrid methods that integrate the strengths of multiple algorithms [10, 11]. Supervised methods handle class imbalance and adapt to shifting data distributions [12], whereas unsupervised methods identify anomalies without labeled data, essential when fraud labels are scarce [13]. Studies integrating autoencoders with automated label-generation techniques for highly imbalanced datasets have demonstrated substantially higher Area Under the Precision–Recall Curve (AUC-PR) performance compared with traditional baseline approaches [14]. Accordingly, an effective fraud detection framework should combine feature engineering, adaptive resampling, and ensemble modeling to ensure high sensitivity to rare fraudulent transactions while maintaining robustness against concept drift [15]. Nonetheless, the complexity of digital transaction dynamics necessitates further investigation to ensure that detection systems can respond to emerging fraud patterns in real time while preserving stable performance on highly imbalanced datasets [16].

Recent studies emphasize the role of resampling techniques integrated with ensemble learning as an effective strategy to address skewed class distributions. Authors in [17] show that combining Synthetic Minority Oversampling Technique (SMOTE) with RF supported by Focal Loss substantially enhances the model's sensitivity to minority-class patterns, leading to improved recall without reducing performance consistency. This finding is reinforced by authors in [18], report that applying SMOTE with a Random Forest classifier significantly improves key performance metrics, including recall, F1-score, and overall detection robustness, when compared to models trained on imbalanced data.

Advanced ensemble architectures have also received considerable attention. Authors in [19] investigate boosting and bagging approaches optimized through hyperparameter tuning and feature selection. Their results indicate that ensemble-based models consistently outperform individual classifiers, particularly in terms of recall, which is crucial for identifying rare fraudulent events. Additional support is provided by authors in [20], who demonstrate that resampling methods such as SMOTE and ADASYN, when used alongside XGBoost or RF, can produce accuracy values approaching 99% and an Area Under the Curve (AUC) of 0.993. These outcomes highlight the advantages of structured ensemble pipelines over single-model strategies. Generative models have also contributed important new insights. Authors in [21] employ ESMOTE–GAN to create more realistic representations of the minority class, resulting in higher true positive rates, although challenges associated with Generative Adversarial Network (GAN) training stability remain unresolved. In the spectrum of classical ML algorithms, authors in [22] reaffirm that RF continues to be the most robust baseline compared with Logistic Regression and Decision Tree, especially when dealing with extreme class imbalance. This strand of evidence is extended by authors in [23], who present an ensemble approach that integrates LSTM with SMOTE–ENN. Their findings demonstrate improved recall without triggering overfitting, illustrating the promising synergy between deep learning and hybrid data-level techniques.

Efforts to strengthen model stability are further illustrated in the study by authors in [24], which reports that SMOTE and SMOTE-Tomek maintain AUC values between 0.985 and 0.992 across different configurations. This consistency suggests that feature-oriented resampling techniques are capable of preserving strong performance even in complex evaluation settings. Similar conclusions are noted by authors in [25] whose hybrid SMOTE, RF, and XGBoost configuration achieves an AUC of 0.993 with an accuracy of 99%. These results underscore the capacity of hybrid ensemble models to sustain high recall across diverse testing scenarios.

A foundational contribution in this field is provided by authors in [26], who compare several classical algorithms including Decision Tree, Logistic Regression, Naïve Bayes, KNN, and RF on a pre-processed financial transaction dataset. Their analysis reveals that RF and KNN exhibit stronger fraud detection capabilities, whereas Logistic Regression and Naïve Bayes show limited sensitivity to minority-class instances. This baseline evidence has shaped the development of more advanced fraud detection frameworks. Further refinement is presented by authors in [27], who demonstrate that applying SMOTE in combination with AdaBoost significantly improves detection sensitivity for extremely rare fraudulent transactions [28]. Their findings confirm the important role of boosting methods in elevating recall under conditions of severe imbalance [29].

Despite these advances, existing approaches continue to exhibit three principal limitations. First, generative methods such as GANs require the concurrent training of generator and discriminator models, rendering them susceptible to convergence instability, challenging hyperparameter optimization, and substantial computational overhead. Second, unduly complex ensemble architectures, exemplified by deep stacking techniques, tend to compromise model interpretability and impede inference speed, both of which are critical for real-time fraud detection systems. Third, model evaluation frequently relies on conventional metrics such as accuracy, precision, recall, or Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which exhibit limited sensitivity to minority-class performance under severe class imbalance.

In response to these limitations, this study proposes SMOVO (SMOTE-based Voting Ensemble), a comprehensive framework that integrates three core components: (1) feature selection via Analysis of Variance (ANOVA) F-test to retain statistically discriminative predictors, (2) minority class oversampling using SMOTE, and (3) a heterogeneous soft-voting ensemble classifier that combines RF, XGBoost, LightGBM, and AdaBoost. This integrated architecture is designed to enhance detection stability, preserve model interpretability, and improve computational efficiency.

Performance is evaluated comprehensively using a suite of metrics sensitive to class imbalance, including recall, F1-score, Matthews Correlation Coefficient (MCC), and Geometric mean (G-mean).

A. Machine Learning Approaches for Fraud Detection

Voting-based ensemble classifiers have demonstrated strong efficacy by integrating multiple algorithms to improve predictive robustness and classification performance [30]. The voting model aggregates the outputs of RF, XGBoost, LightGBM, and AdaBoost to produce more accurate and stable classification decisions [31].

The advantage of this ensemble approach lies in its ability to reduce the variance of individual models, enhance generalization, and strengthen sensitivity to anomalous patterns in transaction data [32]. By combining the strengths of diverse algorithms with different characteristics, the fraud detection system becomes more robust in addressing data imbalance and the complexity of financial transaction features.

B. Comparative Overview of ESMOTE-GAN and the SMOVO Framework

Figure 1 illustrates the end-to-end workflow of the proposed SMOVO framework in comparison with the ESMOTE-GAN system. It includes data preprocessing, class balancing using SMOTE, feature selection via ANOVA F-test, and classification using a heterogeneous soft-voting ensemble composed of RF, XGBoost, LightGBM, and AdaBoost. This design emphasizes the sequential integration of resampling, feature selection, and ensemble learning to address extreme class imbalance in fraud detection. For comparative purposes, this study considers the ESMOTE-GAN system as a representative generative oversampling approach. The ESMOTE-GAN system integrates undersampling, SMOTE, and GAN-based data synthesis before classification using RF. While it can generate richer minority sample variations, this approach suffers from high computational complexity, is prone to training instability due to the characteristics of GANs, offers low interpretability, and its evaluation is limited to AUC and False Alarm Rate, thus providing an incomplete assessment of minority class detection. By contrast, the SMOVO framework implements a simpler and more stable architecture by combining SMOTE, ANOVA-based feature selection, and a heterogeneous voting classifier (RF, XGBoost, LightGBM, AdaBoost). This approach is computationally efficient, easily replicable, and highly interpretable as all components are transparent. Comprehensive evaluation using recall, F1-score, AUC, MCC, and G-mean provides a holistic view of model performance on imbalanced data. Overall, SMOVO offers a more robust and practical alternative to GAN-based solutions while maintaining high sensitivity in fraud detection.

C. Research Gap and Contribution

SMOVO enhances stability, computational efficiency, and detection accuracy in highly imbalanced fraud scenarios, improving upon conventional generative approaches. First, this research integrates the ANOVA F-test as a significance-based feature selection mechanism, capable of rapidly and consistently identifying variables with the highest discriminative power. This strategy sharpens the model's focus on relevant transaction behavior characteristics while reducing complexity without compromising generalization ability.

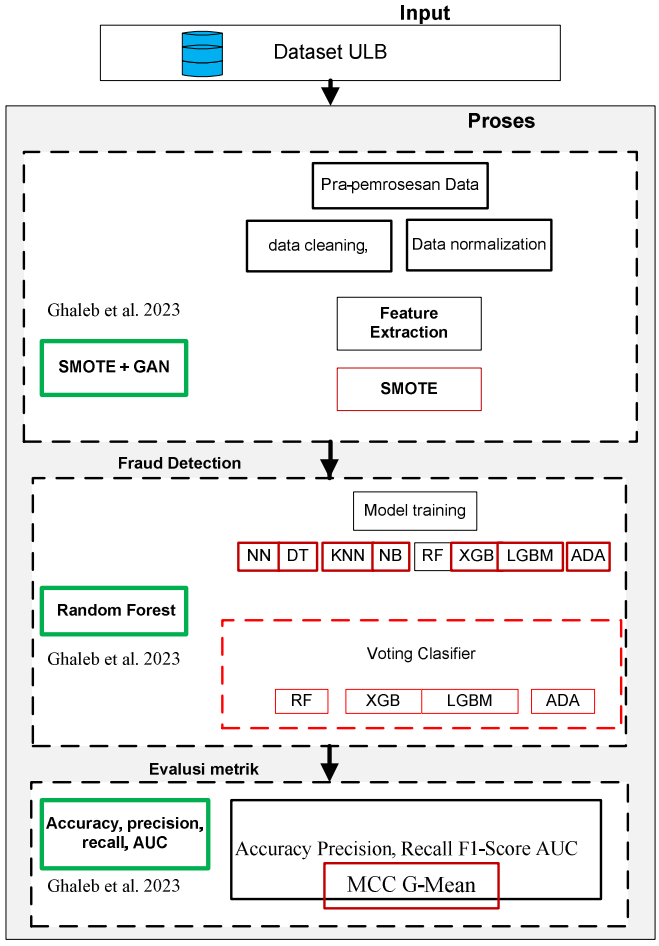

Fig. 1. End-to-end workflow of the proposed SMOVO framework and comparison with ESMOTE-GAN.

Second, the study formulates an adaptive class-balancing process based on SMOTE, which is more stable than GAN-based synthesis techniques. This approach ensures that minority class representations are systematically reinforced without inducing the training variability that often disrupts generative model performance.

Third, the research introduces the integration of a heterogeneous soft voting ensemble combining RF, XGBoost, LightGBM, and AdaBoost to leverage the complementary strengths of each algorithm. This probabilistic combination enhances the reliability of model decisions and increases sensitivity to rare anomalous patterns.

The contribution of this study not only offers a more practical and interpretable alternative to ESMOTE-GAN but also provides a measurable approach that can consistently improve detection performance through multidimensional metric evaluation. The proposed SMOVO framework thus establishes a solid methodological foundation for developing adaptive, efficient fraud detection systems that are ready to be

deployed in large-scale financial data environments. Moreover, unlike GAN-based oversampling approaches that are sensitive to initialization and training dynamics, the deterministic nature of SMOTE, when integrated with ANOVA-based feature selection and ensemble aggregation, contributes to more stable and reproducible performance within the experimental setting of this study; this is evidenced by the consistent improvements in MCC, G-mean, and AUC observed across the conducted evaluations. To provide a comparative context and highlight the positioning of this study's contribution among previous approaches, Table I summarizes several representative studies focused on credit card fraud detection and related domains.

TABLE I. COMPARATIVE SUMMARY OF RECENT STUDIES IN CREDIT CARD TRANSACTION FRAUD DETECTION

Ref	Dataset	Method / Model	Key metrics	Main findings
[17]	EU Credit Card	SMOTE + RF + Focal Loss	Accuracy, recall, AUC	Focal Loss improves minority-class recall when combined with RF and SMOTE
[18]	EU Credit Card	SMOTE + RF	Accuracy, F1-score	SMOTE + RF improves performance in imbalanced fraud detection
[19]	EU Credit Card	DT, LR, NB, KNN, RF	Accuracy, precision, recall, F1-score	RF and KNN show stable performance; LR and NB are less sensitive to minority class
[20]	EU Credit Card	SMOTE / ADASYN, XGB, RF	Accuracy 99.0%, AUC 0.993	Ensemble models outperform single classifiers
[21]	EU Credit Card	ESMOTE-GAN + RF ensemble	Accuracy, precision, F1-score, AUC	GAN improves TPR but requires complex tuning
[22]	EU Credit Card	LR, DT, RF	Accuracy, precision, recall, F1-score, AUC	RF achieves best performance on imbalanced data
[23]	EU Credit Card	LSTM, AdaBoost ensemble, SMOTE-ENN	Accuracy, precision, recall, F1-score, AUC	Neural ensembles improve recall without overfitting
[24]	EU Credit Card	SMOTE / SMOTE-Tomek, ensemble	AUC 0.985–0.992	Feature-based resampling improves stability
[25]	EU Credit Card	SMOTE ensemble (RF, XGB)	Accuracy 99.0%, AUC 0.993	RF–XGB ensemble consistently improves AUC and recall
[26]	EU Credit Card	DT, LR, NB, KNN, RF	Accuracy, precision, recall, F1-score	RF and KNN outperform LR and NB
[27]	EU Credit Card	SMOTE + AdaBoost	Accuracy, AUC, F1-score	Boosting with SMOTE increases fraud sensitivity

II. METHODOLOGY

A. Dataset

This study utilizes a public dataset from the Kaggle platform, comprising credit card transactions by European cardholders in September 2013 [33]. The dataset contains 284,807 transactions over a two-day period, of which 492 are labeled as fraudulent. This highlights the common challenge in fraud detection with imbalanced data, as fraudulent transactions constitute only approximately 0.172% of the total dataset [13]. The target variable in this dataset is Class, a binary indicator where 1 denotes a fraudulent transaction, and 0 denotes a normal transaction [34].

B. Data Preprocessing

The preprocessing stage ensures that the data are of high quality, demonstrate a stable distribution, and maintain an appropriate representation between minority and majority classes. This stage constitutes the first component of the proposed SMOVO pipeline, which is illustrated in Figure 2 and comprises data preprocessing, feature selection, classification, evaluation, and interpretation of results.

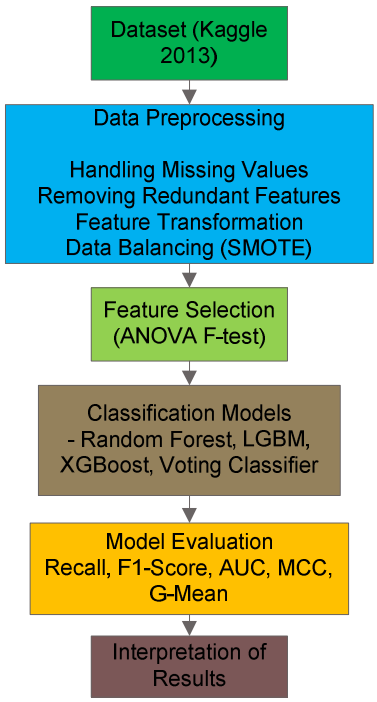

Fig. 2. Proposed SMOVO pipeline.

C. Class Distribution

The class imbalance in the credit card transaction dataset is clearly illustrated Figure 3, which highlights the substantial disparity between non-fraudulent instances (the majority class) and fraudulent activities (the minority class). With fraudulent cases representing only 0.17% of total observations, this extreme imbalance necessitates the use of resampling strategies. Accordingly, the proposed framework incorporates SMOTE to mitigate this disparity and improve the model's predictive performance for the minority class.

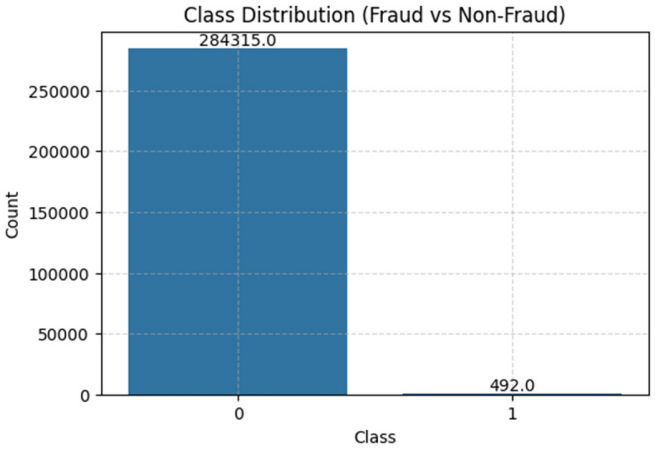

Fig. 3. Histogram of class distribution of the transaction dataset.

D. Feature Selection Using Analysis of Variance F-Test

Feature selection based on the ANOVA F-test evaluates the significance of mean differences among numerical features with respect to a categorical target variable, such as fraudulent versus non-fraudulent classes. The test calculates an F-score as the ratio of between-class variance to within-class variance, with higher F-scores indicating greater discriminative power.

The ANOVA F-test functions as an initial mechanism for identifying transaction attributes that are statistically relevant to fraudulent activity. Features with high F-scores typically capture behavioral patterns that deviate from normal transactions, making them prime candidates for prioritization in predictive model development. This strategy not only improves modeling efficiency but also enhances the interpretability of the fraud detection system by selecting empirically significant variables.

Figure 4 presents the sixteen key features selected based on ANOVA F-Test scores for credit card fraud detection. These features exhibit significant mean differences between fraudulent and non-fraudulent classes, representing the most statistically relevant behavioral indicators. Feature analysis encompasses transaction value, temporal characteristics, device and location consistency, activity intensity, and risk profiling:

- Amount, Transaction_Time, and Avg_Spending capture transaction dynamics.
- Device_ID, Browser_Type, and Login_Location_Change assess the stability of the user's digital identity.
- Geo_Distance and Country_Match, detect location anomalies.
- Frequency_24h and Time_Delta evaluate transaction intensity.
- Card_Limit_Usage, IP_Risk_Score, and Num_Past_Fraud reflect financial and historical risk.
- Merchant_Type, Account_Age, and Transaction_Channel assist in identifying deviations from typical user behavior patterns.

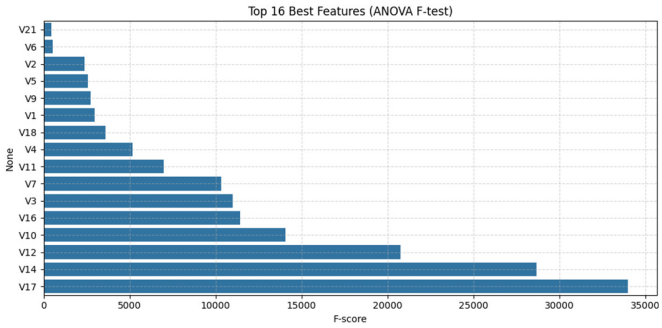

Fig. 4. Key features selected using the ANOVA F-test within the SMOVO framework.

The ANOVA F-test was selected due to its statistical rigor, computational efficiency, and interpretability. This method quantifies the discriminative power of each feature independently of any predictive model. High F-values indicate significant differences between the means of fraudulent and non-fraudulent transactions, identifying features that are strong candidates for classification. It is particularly suitable for numerical features and large-scale datasets, as it enables rapid evaluation without iterative model training while maintaining transparency for regulatory compliance.

E. Evaluation Model

1) XGBoost

XGBoost demonstrated consistent performance in detecting fraudulent patterns within highly imbalanced financial datasets. Leveraging gradient boosting and second-order optimization, XGBoost effectively captures complex non-linear relationships among transaction features [35].

2) LightGBM

LightGBM offers high training efficiency through its leaf-wise growth and histogram-based approaches, accelerating training without compromising accuracy. It has proven effective and stable in detecting credit card fraud despite variations and noise in financial data [36].

3) AdaBoost

AdaBoost operates adaptively by assigning higher weights to misclassified instances. Although sensitive to synthetic data, its performance improves when combined with hybrid resampling techniques, enhancing sensitivity to the minority class in credit card fraud detection [37].

4) Random Forest

RF is a bagging ensemble that produces stable and interpretable predictions while minimizing overfitting. It combines high accuracy with interpretability through feature importance analysis, making it reliable for credit card fraud detection.

5) Voting Classifier (Soft Voting)

The soft voting ensemble combines the strengths of base models such as XGBoost, LightGBM, AdaBoost, and RF to

enhance classification stability and accuracy, leveraging the complementary strengths of bagging and boosting [38].

6) Evaluation Metrics

Model performance was evaluated considering the highly imbalanced class distribution and the dynamic nature of user behavior [39]. Accordingly, five evaluation metrics were utilized to deliver a more balanced and operationally meaningful assessment of model performance [40].

a) Matthews Correlation Coefficient

MCC was selected due to its ability to provide a balanced evaluation across all components of the confusion matrix. Unlike accuracy, which can be easily biased toward the majority class, MCC remains stable even when fraudulent transactions occur in very small proportions. In the context of credit card user behavior, MCC helps ensure that the model can distinguish normal behavioral variations, such as seasonal spending patterns, from anomalous patterns associated with fraudulent activity [41].

b) Geometric Mean

G-mean evaluates the balance between sensitivity and specificity, making it particularly important when user behavior exhibits natural fluctuations that may resemble anomalies. By employing G-mean, the model is required not only to detect fraudulent transactions but also to consistently classify normal transactions. This is crucial, as many users may conduct high-volume transactions over certain periods or make purchases at new locations that remain legitimate [42].

c) Recall

Recall is prioritized because failing to identify fraudulent transactions can lead to significant financial losses. In practice, low-value transactions conducted from previously unused devices often serve as early indicators of account compromise. High recall ensures that the model remains sensitive to such patterns, even when they closely resemble legitimate behavior [25].

d) F1-Score

The F1-score integrates precision and recall, providing a balanced view of model performance in environments characterized by behavioral ambiguity. In practice, users may exhibit 'fraud-like' behavior in certain situations, such as making high-frequency purchases during discount periods or conducting transactions outside their usual geographic location. F1-score ensures that improvements in detection capability do not compromise the accuracy of identifying legitimate transactions [43].

e) Area Under the Receiver Operating Characteristic Curve

AUC-ROC assesses the model's ability to distinguish between fraudulent and non-fraudulent transactions across all decision thresholds, summarizing overall discriminative capability.

On the other hand, accuracy and precision were not employed, as they do not adequately reflect model performance on highly imbalanced data. Accuracy is easily biased toward the majority class, appearing high even when the model fails to identify fraudulent transactions [44]. Precision is also inadequate, as it only evaluates the correctness of fraud predictions without considering the model's ability to capture all fraudulent instances.

F. Experimental Setup and Configuration

To ensure methodological rigor and reproducibility, we designed a structured experimental pipeline. The dataset was partitioned using a stratified 80:20 split for training and testing to preserve the original class distribution, and five-fold cross-validation was applied exclusively to the training set to mitigate overfitting. SMOTE oversampling was configured with k_neighbors = 5 and implemented only within the training folds, whereas all sources of randomness were controlled using random_state = 42 to guarantee consistent and reproducible outcomes.

Model hyperparameters were systematically optimized as follows: RF (n_estimators = 200, max_depth = 20), XGBoost (n_estimators = 200, learning_rate = 0.1, max_depth = 6), LightGBM (n_estimators = 200, num_leaves = 50), and AdaBoost (n_estimators = 100, learning_rate = 1.0).

All experiments were conducted in Python 3.9 using established libraries, including scikit-learn 1.3.0, XGBoost 1.7.0, LightGBM 4.1.0, and imbalanced-learn 0.11.0. This implementation ensured a systematic experimental workflow, verifiable results, and the robustness of the study's empirical findings.

III. RESULTS AND INTERPRETATION

The results confirm that the SMOVO framework delivers superior fraud detection performance compared with conventional approaches and single classifiers. The evaluation considered four widely used algorithms, namely RF, XGBoost, LightGBM, and AdaBoost, each assessed independently, as well as an ensemble configuration implemented through a soft voting classifier, in which final predictions are derived from the probability estimates of the base learners. After the application of SMOTE, performance was measured using recall, F1-score, MCC, AUC-ROC, and G-mean. The complete results are reported in Table II.

TABLE II. MODEL PERFORMANCE WITH AND WITHOUT SMOTE

Model	Recall	F1-score	MCC	AUC	G-mean
RF (no SMOTE)	0.9115	0.8434	0.9224	0.9891	0.9706
Voting classifier (soft, no SMOTE)	0.9423	0.8704	0.8286	0.9121	0.95807
RF	0.9434	0.8600	0.9259	0.9891	0.9712
XGBoost	0.9623	0.8718	0.8754	0.9827	0.9807
LightGBM	0.9245	0.7206	0.7382	0.9480	0.9609
AdaBoost	0.9623	0.0916	0.2110	0.9808	0.9632
Voting classifier (soft)	0.9623	0.8644	0.8686	0.9821	0.9807

The evaluation revealed significant differences in model performance, particularly in handling class imbalance. The application of SMOTE consistently improved ensemble-based classifiers: RF achieved a recall of 0.9434, F1-score of 0.8600, and MCC of 0.9259, whereas the voting classifier attained

higher values with recall 0.9623, F1-score 0.8644, and MCC 0.8686, indicating more consistent identification of minority-class instances.

Among single-model classifiers, XGBoost demonstrated the highest recall (0.9623) and an AUC of 0.9827, although its F1-score (0.8718) and MCC (0.8754) remained moderate. LightGBM maintained a recall of 0.9245, with lower F1-score (0.7206) and MCC (0.7382), reflecting a reduced balance between sensitivity and overall predictive quality. In contrast, AdaBoost exhibited high recall (0.9623) but extremely low F1-score (0.0916) and MCC (0.2110), indicating a substantial proportion of incorrect positive predictions and limited practical applicability.

Overall, the SMOTE-based voting classifier delivered the most balanced performance, achieving high recall (0.9623) while maintaining stable F1-score (0.8644) and MCC (0.8686). These results confirm that combining ensemble methods with class balancing enhances predictive reliability for minority-class instances. Accuracy and precision were not emphasized due to their limitations under severe class imbalance: accuracy is biased toward the majority class, and precision does not account for undetected fraud cases (false negatives). Computational efficiency was also assessed to evaluate the trade-off between predictive performance and processing cost, a critical consideration for fraud detection systems that require frequent model updates in dynamic data environments.

A. Training Time

To compare computational efficiency, the average training time was recorded for each algorithm. The comparative results are summarized in Table III.

TABLE III. AVERAGE TRAINING TIME

Model	Training time (s)
RF	75.3047
XGBoost	1.8225
LightGBM	2.1267
AdaBoost	2.4267
Voting classifier (soft)	82.7677

From a computational efficiency perspective, notable differences were observed in the training times across the models. RF required 75.30 s, whereas the voting classifier (soft) recorded the longest duration at 82.77 s, which aligns with the complexity of aggregating predictions from multiple base learners. In contrast, XGBoost and LightGBM demonstrated substantially higher efficiency, with training times of 1.82 s and 2.13 s, respectively, without compromising predictive performance. These findings highlight the advantage of modern boosting algorithms, which leverage gradient optimization and more effective tree pruning to accelerate the learning process.

Although ensemble training requires longer time, SMOVO inference is highly efficient for real-time deployment. Classifying a new transaction involves a single forward pass through each base model followed by instantaneous soft-voting aggregation, with no re-optimization needed. This efficiency stems from the model architecture: XGBoost and LightGBM are optimized for fast inference, RF supports parallel evaluation, and AdaBoost employs lightweight weak learners. The combination delivers low latency, enabling operational fraud detection with a balance of high accuracy and practical deployability.

B. Data Visualization

The performance of the voting classifier following the application of the SMOTE technique (SMOVO) is visualized through confusion matrices and ROC curves to provide a comprehensive depiction of the model's effectiveness in detecting financial fraud within imbalanced datasets. Additionally, ROC curves and MCC visualizations are presented to offer a more detailed perspective on performance changes for each model after data balancing.

Figure 5 shows the confusion matrix of the RF model after applying SMOTE. The matrix summarizes true positives, true negatives, false positives, and false negatives, illustrating the model's ability to correctly classify both fraudulent and non-fraudulent transactions under class imbalance.

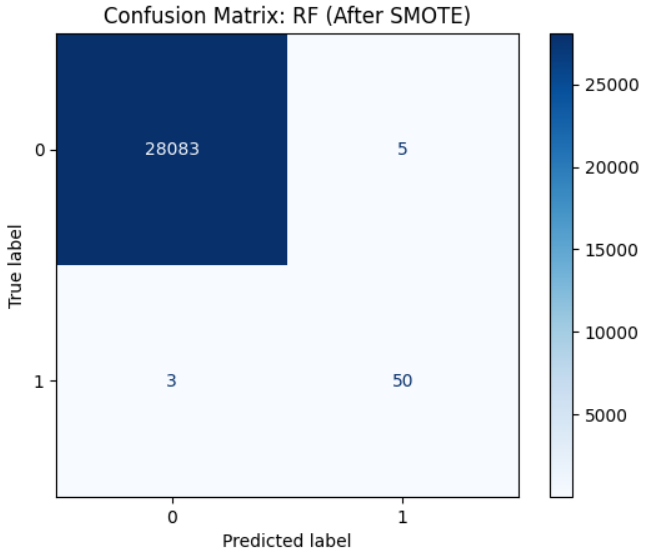

Fig. 5. Confusion matrix of RF after SMOTE.

Figure 6 presents the ROC curve of the RF model after applying SMOTE, depicting the trade-off between true positive rate and false positive rate across different thresholds. The AUC indicates strong discriminative capability in distinguishing fraud from legitimate transactions.

Figure 7 shows the confusion matrix of the soft-voting ensemble (SMOVO) after applying SMOTE. The matrix provides a detailed overview of the model's classification performance, emphasizing its high sensitivity to fraudulent transactions and low misclassification rate for the majority class. The confusion matrix illustrates that the model correctly classified 28,074 non-fraud samples (majority class) and 51 out of 53 fraud samples (minority class), resulting in only 14 false positives and 2 false negatives. This demonstrates the model's high sensitivity toward the minority class while maintaining a low misclassification rate for legitimate transactions, a critical

requirement for operational fraud detection systems. The high number of true positives in the fraud class highlights the model's effectiveness in minimizing undetected fraudulent activities, thereby reducing potential financial losses.

Figure 8 illustrates the ROC curve of the voting classifier (SMOVO) after applying SMOTE. The AUC of 0.9821 indicates highly robust and stable performance in detecting fraudulent transactions, even under extreme class imbalance. The ROC curve's position well above the diagonal random baseline confirms that the model achieves near-optimal predictive accuracy, effectively distinguishing between minority (fraud) and majority (legitimate) classes.

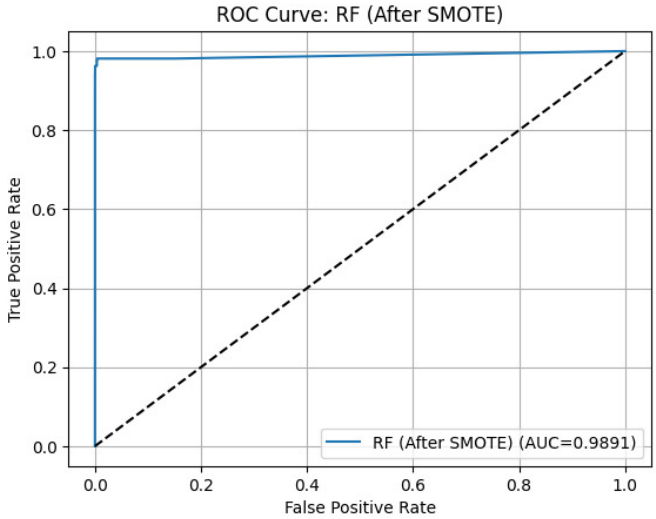

Fig. 6. ROC curve of RF after SMOTE.

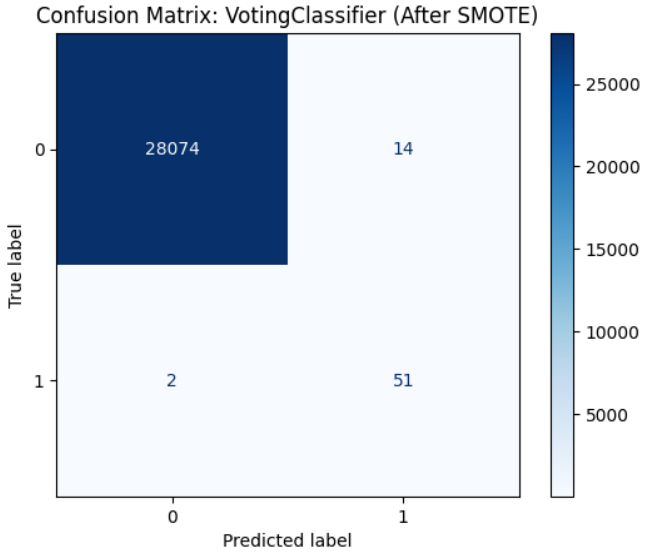

Fig. 7. Confusion matrix of the voting classifier after SMOTE.

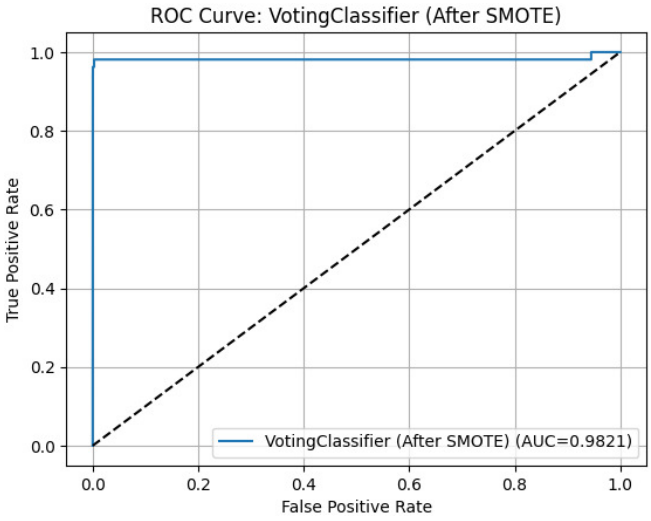

Fig. 8. ROC curve of the voting classifier after SMOTE (SMOVO).

IV. COMPARATIVE EVALUATION

The performance of the SMOVO framework is further analyzed by comparing its classification outcomes with individual algorithms (RF, XGBoost, LightGBM, and AdaBoost) as well as with previously reported approaches. The discussion focuses on recall, F1-score, MCC, AUC, and G-mean, which are particularly informative for datasets with pronounced class imbalance. Table IV summarizes the comparative performance of the SMOVO-based voting classifier alongside RF and the ESMOTE–GAN approach reported by authors in [21].

TABLE IV. COMPARATIVE EVALUATION OF SMOVO AND BASELINE METHODS

Model	Recall	F1-score	MCC	AUC	G-mean	Training time (s)
Voting classifier (SMOVO)	0.9623	0.8644	0.8686	0.9821	0.9807	82.7677
RF (SMOVO)	0.9434	0.9259	0.9259	0.9891	0.9712	75.3047
ESMOTE–GAN [21]	0.93	0.92	–	0.92	–	–

The results demonstrate the effectiveness of integrating SMOTE with a heterogeneous voting ensemble within the SMOVO framework. The SMOVO-based voting classifier achieved the highest recall (0.9623), indicating superior sensitivity in detecting fraudulent transactions, and the highest G-mean (0.9807), reflecting balanced performance between minority (fraud) and majority (legitimate) classes. Concurrently, RF under the SMOVO configuration exhibited the highest F1-score and MCC (both 0.9259) alongside a very high AUC (0.9891), confirming robust discriminative capability across decision thresholds.

Compared with the ESMOTE–GAN approach, SMOVO demonstrates both methodological and practical advantages. ESMOTE–GAN reported competitive recall, F1-score, and AUC values (~0.92), but did not provide MCC, G-mean, or computational time analysis. SMOVO offers a measurable recall improvement of 4.6% and the highest G-mean, indicating

more consistent discrimination between minority and majority classes under extreme imbalance.

Architecturally, SMOVO provides greater stability and transparency than GAN-based solutions. GAN training requires simultaneous optimization of generator and discriminator networks, which introduces convergence instability, increases hyperparameter sensitivity, and imposes substantial computational overhead. By combining SMOTE with a tree-based soft-voting ensemble, SMOVO avoids these complexities while preserving interpretability through explicit contributions from individual base learners, enhancing suitability for operational deployment.

Metric selection reflects the limitations of conventional performance indicators in highly imbalanced datasets. Accuracy and precision are biased toward the majority class and fail to account for the operational cost of undetected fraud. Accordingly, this study emphasizes recall, F1-score, MCC, and G-mean as more reliable measures of detection effectiveness in real-world fraud detection scenarios.

V. CONCLUSION AND FUTURE WORK

This study proposes SMOVO, a Synthetic Minority Oversampling Technique (SMOTE)-based Voting Ensemble framework for credit card fraud detection under extreme class imbalance. SMOVO integrates feature selection via Analysis of Variance (ANOVA) F-test, SMOTE resampling, and a heterogeneous soft-voting ensemble comprising Random Forest (RF), XGBoost, LightGBM, and AdaBoost. Experimental results on a highly imbalanced real-world dataset demonstrate that the proposed SMOVO framework consistently outperforms individual classifiers and conventional baselines when evaluated using imbalance-sensitive metrics. In particular, the RF and the voting classifier within SMOVO achieve concurrent improvements in recall, F1-score, and Matthews Correlation Coefficient (MCC), indicating a balanced trade-off between minority-class sensitivity and overall predictive stability. These findings further confirm that high recall alone is insufficient for reliable fraud detection, as excessive sensitivity may degrade other critical performance measures, particularly F1-score and MCC.

From a computational perspective, ensemble training incurs higher cost than single-model approaches; however, inference remains efficient. Fraud prediction requires only parallel evaluation of pre-trained base learners followed by lightweight soft-voting aggregation, enabling low-latency decision-making suitable for real-time transaction monitoring. While training time has been reported, inference latency and large-scale throughput were not explicitly evaluated and warrant further investigation.

Compared with generative oversampling methods such as ESMOTE-GAN, SMOVO offers clear methodological and practical advantages. Although ESMOTE-GAN can produce more diverse minority samples, it is often associated with high computational complexity, training instability, and limited interpretability. In contrast, SMOVO relies on deterministic, transparent, and easily reproducible components, while providing a more comprehensive evaluation using MCC, G-mean, and Area Under the Curve (AUC). The results confirm that SMOVO maintains high sensitivity with superior overall performance balance.

SMOVO was evaluated using the European Credit Card Fraud Dataset, a standard benchmark for fraud detection under extreme class imbalance. Relying on a single dataset limits the generalizability of the findings, as differences in transaction structures, fraud types, and feature distributions across networks and jurisdictions may affect performance. The results demonstrate that SMOVO is effective within the tested domain, but cross-dataset evaluation using diverse sources is necessary to assess transferability, scalability, and robustness. Financial transaction data are inherently dynamic due to evolving user behavior and adaptive fraud strategies. Although SMOVO performs well on historical data, it does not explicitly address concept drift. Future work will incorporate online and incremental learning, sliding-window retraining, dynamic resampling, cross-dataset validation, inference benchmarking, and production-scale testing to evaluate scalability, generalizability, and comparative performance against generative approaches.

ACKNOWLEDGMENT

The authors wish to express their sincere appreciation to Malang State University (UM) for providing academic guidance, research support, and a constructive scholarly environment that significantly facilitated the completion of this study. The support from UM was instrumental in strengthening the research design and ensuring the overall quality of the work.

The authors also extend their gratitude to the Oriental University of Timor Lorosa'e (UNITAL), Timor-Leste, for their collaboration, institutional support, and valuable contributions throughout the research process. The engagement and cooperation from UNITAL were crucial in enabling the smooth progression and successful completion of this research project.

REFERENCES

[1] M. Barroso and J. Laborda, "Digital transformation and the emergence of the Fintech sector: Systematic literature review," *Digital Business*, vol. 2, no. 2, Jan. 2022, Art. no. 100028, https://doi.org/10.1016/j.digbus.2022.100028.

[2] H. Chitimira, "Overview analysis of the regulation of insider trading in Zimbabwe during the corona virus pandemic," *Journal of Financial Crime*, vol. 30, no. 6, pp. 1499–1516, July 2023, https://doi.org/10.1108/JFC-04-2023-0089.

[3] T. Widiyaningtyas, D. D. Prasetya, and H. W. Herwanto, "Time Loss Function-based Collaborative Filtering in Movie Recommender System," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 6, pp. 1021–1030, Oct. 2023, https://doi.org/10.22266/ijies2023.1231.84.

[4] P. Sood, C. Sharma, S. Nijjer, and S. Sakhuja, "Review the role of artificial intelligence in detecting and preventing financial fraud using natural language processing," *International Journal of System Assurance Engineering and Management*, vol. 14, no. 6, pp. 2120–2135, Dec. 2023, https://doi.org/10.1007/s13198-023-02043-7.

[5] M. A. K. Achakzai and J. Peng, "Detecting financial statement fraud using dynamic ensemble machine learning," *International Review of Financial Analysis*, vol. 89, Oct. 2023, Art. no. 102827, https://doi.org/10.1016/j.irfa.2023.102827.

[6] M. A. Alrasheedi, "Enhancing Fraud Detection in Credit Card Transactions: A Comparative Study of Machine Learning Models,"

Computational Economics, Aug. 2025, https://doi.org/10.1007/s10614-025-11071-3.

[7] J. Nicholls, A. Kuppa, and N.-A. Le-Khac, "Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape," *IEEE Access*, vol. 9, pp. 163965–163986, 2021, https://doi.org/10.1109/ACCESS.2021.3134076.

[8] E. Oztemel and M. Isik, "A Systematic Review of Intelligent Systems and Analytic Applications in Credit Card Fraud Detection," *Applied Sciences*, vol. 15, no. 3, Jan. 2025, Art. no. 1356, https://doi.org/10.3390/app15031356.

[9] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *Journal of Big Data*, vol. 10, no. 1, Oct. 2023, Art. no. 154, https://doi.org/10.1186/s40537-023-00821-5.

[10] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022, https://doi.org/10.1109/ACCESS.2022.3166891.

[11] P. Gupta, A. Varshney, M. R. Khan, R. Ahmed, M. Shuaib, and S. Alam, "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques," *Procedia Computer Science*, vol. 218, pp. 2575–2584, Jan. 2023, https://doi.org/10.1016/j.procs.2023.01.231.

[12] S. Sendari, I. A. E. Zaeni, D. C. Lestari, and H. P. Hariyadi, "Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm," *Knowledge Engineering and Data Science*, vol. 3, no. 1, pp. 50–59, June 2020, https://doi.org/10.17977/um018v3i12020p50-59.

[13] W. Yundong, A. Zhulev, and O. G. Ahmed, "Credit Card Fraud Identification using Logistic Regression and Random Forest," *Wasit Journal of Computer and Mathematics Science*, vol. 2, no. 3, pp. 1–8, Sept. 2023, https://doi.org/10.31185/wjcms.184.

[14] R. K. L. Kennedy, F. Villanustre, T. M. Khoshgoftaar, and Z. Salekshahrezaee, "Synthesizing class labels for highly imbalanced credit card fraud detection data," *Journal of Big Data*, vol. 11, no. 1, Mar. 2024, Art. no. 38, https://doi.org/10.1186/s40537-024-00897-7.

[15] M. E. Lokanan, "Predicting mobile money transaction fraud using machine learning algorithms," *Applied AI Letters*, vol. 4, no. 2, Apr. 2023, Art. no. e85, https://doi.org/10.1002/ail2.85.

[16] A. Hanae, B. Abdellah, E. Saida, and G. Youssef, "End-to-End Real-time Architecture for Fraud Detection in Online Digital Transactions," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 749–757, June 2023, https://doi.org/10.14569/IJACSA.2023.0140680.

[17] T. Albalawi and S. Dardouri, "Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation," *Frontiers in Artificial Intelligence*, vol. 8, Oct. 2025, Art. no. 1643292, https://doi.org/10.3389/frai.2025.1643292.

[18] Y. Wang, "Fraud detection based on FS-SMOTE model for credit card," *Highlights in Science, Engineering and Technology*, vol. 70, pp. 316–323, Nov. 2023, https://doi.org/10.54097/hset.v70i.12479.

[19] A.-A. Al-Maari, M. Abdulnabi, Y. Nathan, A. Ali, U. Ali, and M. Khan, "Optimized Credit Card Fraud Detection Leveraging Ensemble Machine Learning Methods," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22287–22294, June 2025, https://doi.org/10.48084/etasr.10287.

[20] Z. Y. Lim, Y. H. Pang, K. Z. B. Kamarudin, S. Y. Ooi, and F. S. Hiew, "Bayesian optimization driven strategy for detecting credit card fraud with Extremely Randomized Trees," *MethodsX*, vol. 13, Dec. 2024, Art. no. 103055, https://doi.org/10.1016/j.mex.2024.103055.

[21] F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. Al-Hadhrami, "Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection," *IEEE Access*, vol. 11, pp. 89694–89710, 2023, https://doi.org/10.1109/ACCESS.2023.3306621.

[22] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions," *Decision Analytics Journal*, vol. 6, Mar. 2023, Art. no. 100163, https://doi.org/10.1016/j.dajour.2023.100163.

[23] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022, https://doi.org/10.1109/ACCESS.2022.3148298.

[24] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *Journal of Big Data*, vol. 10, no. 1, Jan. 2023, Art. no. 6, https://doi.org/10.1186/s40537-023-00684-w.

[25] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach," *Big Data and Cognitive Computing*, vol. 8, no. 1, Jan. 2024, Art. no. 6, https://doi.org/10.3390/bdcc8010006.

[26] P. Manorom, U. Detthamrong, and W. Chansanam, "Comparative Assessment of Fraudulent Financial Transactions using the Machine Learning Algorithms Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Random Forest," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15676–15680, Aug. 2024, https://doi.org/10.48084/etasr.7774.

[27] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, https://doi.org/10.1109/ACCESS.2021.3134330.

[28] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1310–1318, Sept. 2024, https://doi.org/10.62527/joiv.8.3.2283.

[29] T. Widiyaningtyas, I. Hidayah, and T. B. Adji, "Recommendation Algorithm Using Clustering-Based UPCSim (CB-UPCSim)," *Computers*, vol. 10, no. 10, Oct. 2021, Art. no. 123, https://doi.org/10.3390/computers10100123.

[30] M. M. Ismail and M. A. Haq, "Enhancing Enterprise Financial Fraud Detection Using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14854–14861, Aug. 2024, https://doi.org/10.48084/etasr.7437.

[31] M. A. Mim, N. Majadi, and P. Mazumder, "A soft voting ensemble learning approach for credit card fraud detection," *Heliyon*, vol. 10, no. 3, Feb. 2024, Art. no. e25466, https://doi.org/10.1016/j.heliyon.2024.e25466.

[32] A. R. Manga', A. N. Handayani, H. W. Herwanto, R. A. Asmara, Y. I. Sulistya, and K. Kasmira, "Analysis of the Ensemble Method Classifier's Performance on Handwritten Arabic Characters Dataset," *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 186–192, Apr. 2023, https://doi.org/10.33096/ilkom.v15i1.1357.186-192.

[33] "Credit Card Fraud Detection." Kaggle, 2013, [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

[34] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "SCARFF: A scalable framework for streaming credit card fraud detection with spark," *Information Fusion*, vol. 41, pp. 182–194, May 2018, https://doi.org/10.1016/j.inffus.2017.09.005.

[35] O. Michael, O. Christopher, A. Maduabuchuku, N. O. Miracle, E.-O. chinenye Love, and O. C. Akachukwu, "Fraud Detection in Financial Transactions Using Ensemble Machine Learning Models," *International Journal of Research and Scientific Innovation*, vol. 12, no. 13, pp. 27–36, Oct. 2025, https://doi.org/10.51244/IJRSI.2025.1213CS003.

[36] C. Iscan, O. Kumas, F. P. Akbulut, and A. Akbulut, "Wallet-Based Transaction Fraud Prevention Through LightGBM With the Focus on Minimizing False Alarms," *IEEE Access*, vol. 11, pp. 131465–131474, 2023, https://doi.org/10.1109/ACCESS.2023.3321666.

[37] M. Alamri and M. Ykhlef, "Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data," *IEEE Access*, vol. 12, pp. 14050–14060, 2024, https://doi.org/10.1109/ACCESS.2024.3357091.

[38] A. A. Alhashmi, A. M. Alashjaee, A. A. Darem, A. F. Alanazi, and R. Effghi, "An Ensemble-based Fraud Detection Model for Financial Transaction Cyber Threat Classification and Countermeasures,"

Engineering, Technology & Applied Science Research, vol. 13, no. 6, pp. 12433–12439, Dec. 2023, https://doi.org/10.48084/etasr.6401.

[39] W. Rahayu *et al.*, "Synthetic Minority Oversampling Technique (SMOTE) for Boosting the Accuracy of C4.5 Algorithm Model," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 3, no. 3, pp. 624–630, June 2024, https://doi.org/10.59934/jaiea.v3i3.469.

[40] O. A. Adepegba, S. A. Akinboro, S. O. Olabiyisi, O. G. Lala, A. A. Onamade, and A. A. Aroyehun, "Credit Card Fraud Detection Using Voter Ensemble Technique (VET)," *Computing, Information Systems, Development Informatics & Allied Research Journa*, vol. 13, no. 2, pp. 15–20, June 2022, https://doi.org/10.22624/AIMS/CISDI/V13N2P3.

[41] P. Purnawansyah *et al.*, "Congestion Predictive Modelling on Network Dataset Using Ensemble Deep Learning," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1597–1613, Oct. 2024, https://doi.org/10.47738/jads.v5i4.333.

[42] D. Breskuvienė and G. Dzemyda, "Enhancing credit card fraud detection: highly imbalanced data case," *Journal of Big Data*, vol. 11, no. 1, Dec. 2024, Art. no. 182, https://doi.org/10.1186/s40537-024-01059-5.

[43] M. R. Baker, Z. N. Mahmood, and E. H. Shaker, "Ensemble Learning with Supervised Machine Learning Models to Predict Credit Card Fraud Transactions," *Revue d'Intelligence Artificielle*, vol. 36, no. 4, pp. 509–518, Aug. 2022, https://doi.org/10.18280/ria.360401.

[44] Y. Xie, A. Li, L. Gao, and Z. Liu, "A Heterogeneous Ensemble Learning Model Based on Data Distribution for Credit Card Fraud Detection," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, July 2021, Art. no. 2531210, https://doi.org/10.1155/2021/2531210.