Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 221–230
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46308

Evaluating The Effectiveness of Augmentation and Classifier Algorithms for Fraud Detection: Comparing CGAN and SMOTE with Random Forest and XGBoost

Sarmini[1*], Sunardi[2], Abdul Fadlil[3]

***Abstract*—Fraud detection in imbalanced datasets, where fraudulent transactions represent a small fraction of total data, presents a major challenge for machine learning models. Traditional classifiers often perform poorly in such scenarios due to their bias toward the majority class. This study investigates the effectiveness of two data augmentation techniques, Synthetic Minority Over-sampling Technique (SMOTE) and Conditional Generative Adversarial Networks (CGAN) in improving fraud detection performance. Both methods are applied to balance the dataset, and their impact is evaluated using two classifiers: Random Forest (RF) and XGBoost. The models are tested across three versions of the dataset: the original imbalanced data, the SMOTE-augmented data, and the CGAN-augmented data. Evaluation metrics include accuracy, precision, recall, F1 score, and ROC-AUC. Results indicate that both augmentation techniques enhance the models' ability to detect fraudulent transactions compared to the original dataset. Notably, CGAN outperforms SMOTE in terms of recall and F1 score, suggesting its ability to generate more diverse and realistic synthetic samples. While SMOTE creates new samples through interpolation, CGAN uses an adversarial process involving a generator and a discriminator, resulting in more complex data representations. The study also finds that XGBoost combined with CGAN yields the highest performance, effectively capturing intricate fraud patterns. In contrast, SMOTE, though beneficial, shows limited capacity in improving recall. This research highlights the importance of advanced augmentation techniques like CGAN in addressing class imbalance and improving fraud detection systems. It also opens pathways for future exploration of deep learning-based augmentation and classification methods in fraud detection.**

***Index Terms*—Fraud detection, SMOTE, CGAN, data augmentation, imbalanced datasets, random forest, XGBoost.**

Received: 3 May 2025; Revised: 16 June 2025; Accepted: 1 July 2025.
*Corresponding author

[1]Sarmini, Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia (e-mail: 2437083013@webmail.uad.c.id).
[2]Sunardi, Department of Electrical Engineering, Universitas Ahmad Daahlan, Yogyakarta, Indonesia (e-mail: sunardi@mti.uad.ac.id).
[3]Abdul Fadlil, Department of Electrical Engineering, Universitas Ahmad Daahlan, Yogyakarta, Indonesia (e-mail: fadlil@mti.uad.ac.id).

I. INTRODUCTION

Fraud detection in industries such as e-commerce and finance remains a critical challenge due to the extreme class imbalance: legitimate transactions often outnumber fraudulent ones by several orders of magnitude. The imbalance, where legitimate transactions vastly outnumber fraudulent ones, makes it challenging for conventional machine learning models to effectively detect fraud. These models often favor the dominant class, leading to false negatives, which can have severe consequences in high-stakes sectors, including financial losses and regulatory issues [1], [2]. Despite extensive research on imbalance-handling techniques, previous studies have not demonstrated how advanced data-driven augmentation strategies perform when transferred from synthetic training sets to real-world fraud scenarios.

This study aims to systematically compare three dataset scenarios the original imbalanced data, a SMOTE-augmented dataset, and a CGAN-generated dataset across Random Forest (RF) and XGBoost classifiers, evaluating their ability to generalize to unseen fraud patterns using ROC-AUC, precision-recall curves, accuracy, and F1-score. We employ SMOTE and conditional GAN (CGAN) to tackle class imbalance. SMOTE synthesizes minority-class samples via interpolation between existing fraud cases, which improves class balance but can oversimplify fraud patterns [3], [4]. In contrast, CGAN uses a generator–discriminator framework to learn complex fraud distributions and produce more realistic synthetic examples, albeit with higher computational cost and tuning complexity [5], [6], [7]. To enhance model discrimination, we derive three categories of features: transaction irregularities (e.g., unusual time gaps or amount spikes), behavioral patterns (e.g., rapid sequences of logins or purchases), and relational features mapping relationships among accounts, IP addresses, and payment instruments. These engineered attributes have been shown to boost precision and recall by highlighting subtle indicators of fraud [8], [9].

This study leverage RF and XGBoost as primary modeling algorithms. RF mitigates variance by aggregating the

predictions of multiple decision trees, offering robustness to noisy features yet requiring careful tuning on imbalanced data [10]. XGBoost, a gradient-boosted tree method, sequentially corrects errors from prior trees to achieve high accuracy and flexibility, though it is sensitive to hyperparameter settings and may overfit without proper regularization [11].

Innovative data augmentation approaches, such as Generative Adversarial Networks (GANs) and adaptive techniques like ADASYN, mark a shift by adjusting to data patterns and generating realistic fraud instances without the overfitting risks associated with conventional methods [12], [13]. When combined with ensemble classifiers, these innovations offer a comprehensive solution to fraud detection, bridging the gap between research and practical application [14], [15].

False negatives in credit card fraud detection pose significant economic risks, often resulting in substantial monetary losses for both credit card issuers and merchants. While exact figures can vary, it is estimated that billions of dollars are lost annually to credit card fraud, significantly affecting e-commerce platforms and financial institutions [16], [17]. Moreover, a single undetected fraud incident can result in chargebacks that may reach substantial amounts, though the specific figure of $250,000 in one day lacks direct empirical support in the reviewed literature. It has been shown that the necessity for efficient fraud detection systems is critical, as traditional methods can struggle with minimizing false negatives.

Research indicates that advanced machine learning algorithms enhance detection accuracy significantly, thereby protecting the interests of stakeholders and reducing costs associated with fraud [18], [19], [20]. The societal implications of increased fraud are profound, as consumer trust and loyalty can diminish, necessitating continuous innovation in fraud prevention strategies to safeguard both consumers and businesses [21], [22].

II. Research Method

Figure 1 shows the methodology, starting with data preprocessing and dividing the dataset into original, SMOTE-augmented, and CGAN-augmented categories. Models (RF and XGBoost) are trained and tested, with performance evaluated on key metrics and results compared.

A. Data Collection

The dataset, sourced from a multinational company and provided by the authors of [23], contains 297,715 transaction records with 37 features, including transaction amount, account balance, user behavior, and metadata like payment details, shown in Table 1. These features were chosen for their relevance to detecting e-commerce fraud.

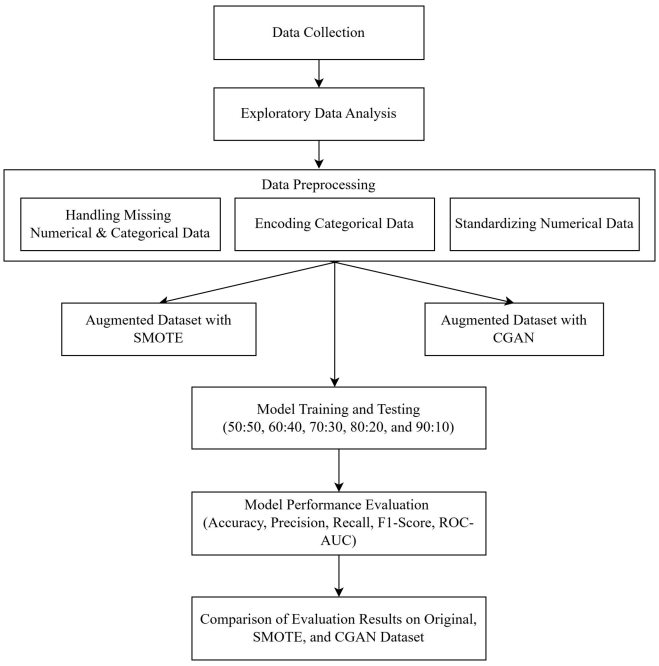

Fig. 1. Research method.

Table 1.
Features of Dataset

No	Column
1	acct_cre_dt
2	NetLoss(is fraud)
3	Account balance
4	Average monthly payment
5	Account receipt balance
6	Cryptocurrency trading
7	Last transaction anomaly
8	Account status 1 (Closed)
9	Account status 2 (Locked)
10	Account status 3 (Restricted)
11	Hosting Server IP
12	Email used for account addition
13	Address used for account addition
14	Credit card used for addition
15	Bank card used for addition
16	Transactions post account creation
17	Country of first credit card used
18	Country of last credit card used
19	Country of first bank account used
20	Country of last bank account used
21	Buyer-seller collusion indicator
22	Reason for user complaint
23	Authorization info used in transactions
24	Goods mismatch indicator
25	Non-receipt of goods indicator
26	Unauthorized access indicator
27	Account-associated IP address
28	Account-associated shipping address
29	Account-associated email
30	Account creation date correlation
31	Account-associated username
32	Credit card associations
33	Account-associated mobile number
34	First credit card affiliation
35	Last credit card linked
36	First bank account linked
37	Last bank account linked

The dataset reveals a substantial imbalanced class, with fraudulent activities constituting only 3.1% of the total transactions, as shown in Table 2.

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 221–230
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46308

Table 2.
Class Distribution in Original Dataset

NetLoss(is fraud)	Count	Percentage(%)
0 (Non-Fraudulent)	288,486	96.90
1 (Fraudulent)	9,229	3.10
Total	297,715	100

Class imbalance is a challenge in machine learning, as models often prioritize the majority class, reducing fraud detection effectiveness. This study addresses the issue using data augmentation and ensemble learning. Key numerical features, like Average monthly payment and Last transaction anomaly, require normalization due to high skewness. Categorical features, such as Account exception status, are crucial for detecting fraud. The dataset also has missing values and duplicates, which need proper handling. Correlation analysis reveals relationships, guiding feature selection and engineering.

B. Data Preprocessing

Preprocessing is vital for ensuring that the e-commerce fraud dataset is clean and suitable for machine learning. In this study, missing values are addressed using imputation techniques: numerical features like Average monthly payment and Balance of receipts in accounts are imputed with the median to reduce the impact of outliers, while categorical features like Exception Status of Accounts 1 Closed and User name association are imputed with the mode. One-hot encoding is applied to categorical variables to convert them into binary vectors, avoiding multicollinearity by dropping the first category. The dataset has a 3.1% fraud rate, and SMOTE is used to balance the classes, improving fraud detection. Numerical features are standardized with z-score normalization to ensure equal contribution.

C. Data Augmentation

This study addresses class imbalance in the e-commerce fraud dataset using two data augmentation techniques: SMOTE and CGAN. SMOTE, a widely used oversampling method, creates synthetic fraudulent transactions by interpolating between instances of the minority class (fraudulent transactions) and their nearest neighbors, ensuring a balanced dataset. CGAN, a deep learning-based approach, uses a generator to create synthetic fraudulent transactions, while a discriminator evaluates their authenticity. The augmented datasets from both methods are compared to the original imbalanced dataset, and the model performance is evaluated using precision, recall, F1-score, and ROC-AUC for detecting fraudulent transactions.

D. Experimental Setup

All experiments were conducted within a Python (v3.12.2) environment to ensure reproducibility. The primary libraries used included scikit-learn (v1.4.2) for model evaluation, imbalanced-learn (v0.12.3) for the SMOTE implementation, and xgboost (v2.1.1) for the classifier. The CGAN model was built using TensorFlow (v2.16.2). The experiments were run on a hardware setup consisting of an Apple M1 chip (8-core CPU and 8-core GPU) with 16 GB of RAM. On this machine, the CGAN model was trained for 100 epochs with a batch size of 64, a process which took approximately 4 hours to complete.

E. Model Training and Evaluation

Model training and evaluation are key steps in this study, focusing on how well RF and XGBoost address class imbalance in e-commerce fraud detection. Both classifiers are chosen for their robustness with imbalanced datasets. Performance is evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. RF, an ensemble method, combines decision trees to prevent overfitting and handles complex datasets well. XGBoost, a gradient-boosting algorithm, builds trees sequentially to correct errors. Hyperparameter tuning for both models is crucial to optimizing performance and improving fraud detection accuracy.

For each experiment, the dataset was divided into training and testing sets using various ratios (50:50, 60:40, 70:30, 80:20, and 90:10). A stratified random sampling technique was applied during each split. This ensures that the proportion of fraudulent transactions (the minority class) was identical in both the training and testing sets, preventing sampling bias and ensuring a fair evaluation across all configurations.

Various metrics are used to assess the performance of each model, with accuracy defined as the proportion of correctly classified instances:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

In imbalanced datasets, accuracy can be misleading, as it favors the majority class. For instance, a model predicting all transactions as non-fraudulent may achieve high accuracy but fail to detect fraud. Precision, which measures the proportion of true positives out of all positive predictions, is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Precision measures how well a model avoids false positives, which is crucial for fraud detection, as misclassifying legitimate transactions can cause problems. A high precision score indicates the model's effectiveness in predicting fraud. Recall, or sensitivity, quantifies the proportion of true positives identified and is calculated as:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall is vital for fraud detection, reflecting the model's ability to identify fraudulent transactions. A high recall score minimizes undetected fraud and financial losses, but it often

increases false positives, highlighting the precision-recall trade-off. The F1-score, the harmonic mean of precision and recall, balances both metrics and is calculated as:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

The receiver operating characteristic (ROC) curve evaluates a model's ability to distinguish between classes across various thresholds, plotting the true positive rate against the false positive rate. The area under the curve (AUC) represents performance, with 1 being perfect and 0.5 indicating random guessing. ROC-AUC is particularly useful for imbalanced datasets. Metrics like precision, recall, and F1-score, adjusted for class imbalance, are calculated using sklearn.metrics. To improve robustness, 5-fold cross-validation is used, with models trained and evaluated using sklearn and xgboost. Results are stored in CSV files, and trained models are saved with joblib for future use.

To statistically validate the comparison between models, a one-way analysis of variance (ANOVA) followed by a Tukey post-hoc test could be performed on the F1-scores obtained from different configurations. This approach would allow us to determine whether the observed performance differences (e.g., between CGAN and SMOTE) are statistically significant (e.g., $p < 0.05$). However, due to the scope of this study, formal significance testing was not performed and is acknowledged as an area for future work.

III. Result

A. Result of the Original Dataset

Models trained on the original, imbalanced dataset demonstrate the classic challenge of fraud detection: while achieving high accuracy (≈97.4%), their practical effectiveness was poor. As detailed in Table 3, both RF and XGBoost classifiers produced very low recall scores (18-24%), indicating that the vast majority of fraudulent transactions were missed. Despite strong ROC-AUC scores (≈0.90), the low F1-scores (0.31 for RF and 0.35 for XGBoost) confirm that high accuracy is a misleading metric in this context and that without intervention, these models are unsuitable for reliable fraud detection.

Table 3.
Performance Metrics for Classifiers on the Original Dataset

Classifier	Train:Test	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	50:50	0.9735	0.8137	0.1883	0.3058	0.8986
RF	60:40	0.9733	0.7931	0.1858	0.3011	0.9003
RF	70:30	0.9735	0.8124	0.1892	0.3070	0.9017
RF	80:20	0.9737	0.8211	0.1939	0.3138	0.8999
RF	90:10	0.9735	0.8054	0.1928	0.3112	0.8978
XGBoost	50:50	0.9734	0.7150	0.2360	0.3548	0.8979
XGBoost	60:40	0.9733	0.7246	0.2224	0.3403	0.8999
XGBoost	70:30	0.9734	0.7064	0.2416	0.3601	0.9018
XGBoost	80:20	0.9738	0.7487	0.2324	0.3547	0.9025
XGBoost	90:10	0.9737	0.7414	0.2329	0.3545	0.9000

To visually assess the models' ability to distinguish between classes, Figure 2 presents ROC curves for both RF (solid lines) and XGBoost (dashed lines) across all five train-test splits.

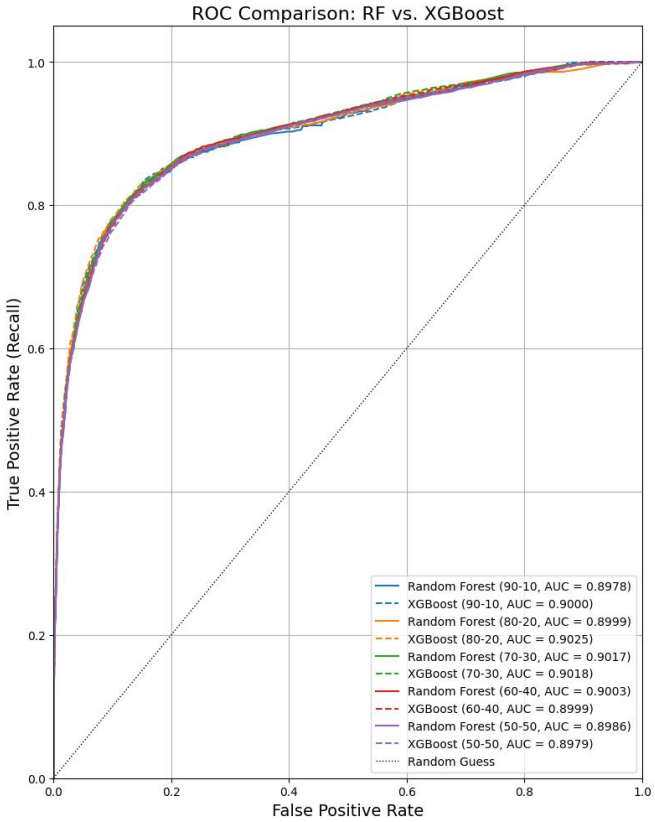

Fig. 2. ROC curves of original dataset.

B. Result of the SMOTE Augmented Dataset

To address the class imbalance, the SMOTE technique was used to generate synthetic fraudulent transactions, creating a perfectly balanced dataset of 576,972 rows. This new dataset features an equal 50/50 split between fraudulent and non-fraudulent instances, as detailed in Table 4. By significantly increasing the representation of the minority class, this augmented dataset provides a more effective training sample for the classifiers.

Table 4.
Class Distribution in SMOTE Augmented Dataset

NetLoss(is fraud)	Count	Percentage (%)
0 (Non-Fraudulent)	288,486	50
1 (Fraudulent)	288,486	50
Total	576,972	100

To test real-world generalization, models were trained on data augmented by SMOTE and then evaluated on the original, imbalanced test set (Table 5). While this approach significantly improved recall to over 77% for both RF and XGBoost, it caused precision to drop dramatically to around 16%. The resulting low F1-scores (≈0.27) highlight a critical trade-off: while the models learned to identify more fraud, they did so at

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 221–230
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46308

the cost of producing an unacceptably high number of false positives, rendering them impractical for production use without further refinement.

Table 5.
Performance Metrics of Training on SMOTE-Augmented Dataset and Testing on Original Dataset

Classifier	Train:Test	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	50:50	0.8665	0.1597	0.7753	0.2648	0.8896
RF	60:40	0.8654	0.1596	0.7833	0.2651	0.8921
RF	70:30	0.8658	0.1606	0.7880	0.2669	0.8954
RF	80:20	0.8652	0.1606	0.7925	0.2671	0.8932
RF	90:10	0.8629	0.1578	0.7887	0.2630	0.8865
XGBoost	50:50	0.8714	0.1624	0.7571	0.2674	0.8709
XGBoost	60:40	0.8694	0.1618	0.7690	0.2674	0.8786
XGBoost	70:30	0.8720	0.1658	0.7754	0.2731	0.8814
XGBoost	80:20	0.8750	0.1692	0.7752	0.2777	0.8856
XGBoost	90:10	0.8668	0.1618	0.7887	0.2685	0.8799

In a controlled environment where both training and testing were performed on the SMOTE-augmented dataset, the models showed significantly improved and more balanced performance. As detailed in Table 6, the XGBoost classifier outperformed RF, achieving a strong F1-score of approximately 0.85, with precision and recall balanced at around 87% and 82%, respectively. The high ROC-AUC score (≈0.93) for XGBoost further confirms that models trained and tested on a balanced SMOTE dataset can effectively differentiate between fraudulent and non-fraudulent transactions.

Table 6.
Performance Metrics of Training and Testing Both on SMOTE-Augmented Dataset

Classifier	Train:Test	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	50:50	0.8665	0.1597	0.7753	0.2648	0.8896
RF	60:40	0.8654	0.1596	0.7833	0.2651	0.8921
RF	70:30	0.8658	0.1606	0.7880	0.2669	0.8954
RF	80:20	0.8652	0.1606	0.7925	0.2671	0.8939
RF	90:10	0.8629	0.1578	0.7887	0.2630	0.8865
XGBoost	50:50	0.8714	0.1624	0.7571	0.2674	0.8709
XGBoost	60:40	0.8694	0.1618	0.7690	0.2674	0.8786
XGBoost	70:30	0.8720	0.1658	0.7754	0.2731	0.8814
XGBoost	80:20	0.8750	0.1692	0.7752	0.2777	0.8856
XGBoost	90:10	0.8668	0.1618	0.7887	0.2685	0.8799

Figure 3 visually summarizes the performance of both pipelines across all train-test splits. The ROC curves for the SMOTE-Ori pipeline (blue and cyan lines) consistently show AUC scores around 0.90. The curves for the SMOTE-SMOTE pipeline (red and magenta lines) show a slight but consistent improvement, with AUC scores reaching up to 0.92.

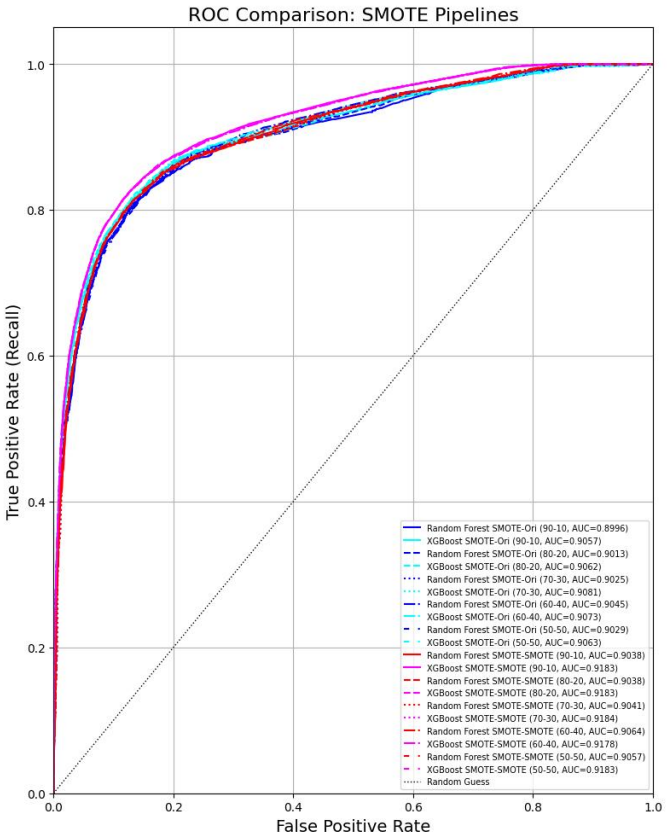

Fig. 3. ROC curves of SMOTE-augmented dataset.

C. *Result of the CGAN Augmented Dataset*

The CGAN-augmented dataset was created using a more advanced generative adversarial framework to address class imbalance. This process resulted in a perfectly balanced dataset of 576,972 rows, with an equal 50/50 split between 288,486 fraudulent and non-fraudulent transactions, as detailed in Table 7. This augmentation provides a larger and more diverse set of fraudulent examples, intended to enhance the classifiers' detection capabilities.

Table 7.
Class Distribution in CGAN Augmented Dataset

NetLoss(is fraud)	Count	Percentage (%)
0 (Non-Fraudulent)	288,486	50
1 (Fraudulent)	288,486	50
Total	576,972	100

When training was performed using the CGAN-augmented dataset and testing on the original dataset, significant issues arose in detecting fraudulent transactions. Both RF and XGBoost models achieved high accuracy, consistently near 97%, suggesting that they performed well in predicting non-fraudulent transactions. Their performance in detecting fraud was poor, with all metrics (precision, recall, F1-score, ROC-AUC) showing zero for fraudulent transactions, indicating failure to identify fraud.

This poor performance suggests that the models may have overfitted to the synthetic patterns in the CGAN-augmented dataset, as these fraudulent examples did not perfectly represent real-world fraud scenarios. The models likely struggled to generalize to the original, imbalanced test data, leading to their failure in detecting fraud. While training on the CGAN-augmented dataset improved accuracy for non-fraudulent transactions, the models severely underperformed in identifying fraud. This highlights the challenges of using synthetic data augmentation techniques like CGAN for fraud detection, necessitating further model calibration or tuning to improve fraud detection on real data. The results of training on the CGAN-augmented dataset and testing on the original dataset are summarized in Table 8.

Table 8.
Performance Metrics of Training on CGAN-Augmented Dataset and Testing on Original Dataset

Classifier	Train:Test	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	50:50	0.9690	0.0000	0.0000	0.0000	0.5005
RF	60:40	0.9690	0.0000	0.0000	0.0000	0.5003
RF	70:30	0.9690	0.0000	0.0000	0.0000	0.5076
RF	80:20	0.9690	0.0000	0.0000	0.0000	0.5000
RF	90:10	0.9690	0.0000	0.0000	0.0000	0.5005
XGBoost	50:50	0.9690	0.0000	0.0000	0.0000	0.7189
XGBoost	60:40	0.9690	0.0000	0.0000	0.0000	0.7180
XGBoost	70:30	0.9690	0.0000	0.0000	0.0000	0.7057
XGBoost	80:20	0.9690	0.0000	0.0000	0.0000	0.4828
XGBoost	90:10	0.9690	0.0000	0.0000	0.0000	0.5534

When both training and testing were performed using the CGAN-augmented dataset, the performance of both the RF and XGBoost models was promising, with consistently high accuracy scores ranging from 95.2% to 95.3%. These results indicate the models were well-calibrated to predict non-fraudulent transactions. However, the key focus in fraud detection is the ability to identify fraudulent transactions, and both models excelled in this area. They achieved precision values around 93.7% to 93.9%, recall values between 96.8% and 97.0%, and F1-scores between 0.95 and 0.96, demonstrating strong performance in detecting fraud. The ROC-AUC scores, ranging from 0.967 to 0.970, demonstrate the models' ability to distinguish fraudulent from non-fraudulent transactions, highlighting the effectiveness of CGAN-based data augmentation in improving fraud detection and reducing false negatives, as shown in the high precision, recall, and ROC-AUC scores in Table 9.

Table 9.
Performance Metrics of Training and Testing Both on CGAN-Augmented Dataset

Classifier	Train:Test	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	50:50	0.9531	0.9388	0.9693	0.9538	0.9679
RF	60:40	0.9531	0.9389	0.9693	0.9539	0.9682
RF	70:30	0.9529	0.9384	0.9696	0.9537	0.9677
RF	80:20	0.9523	0.9377	0.9689	0.9530	0.9675
RF	90:10	0.9522	0.9377	0.9688	0.9527	0.9675
XGBoost	50:50	0.9531	0.9390	0.9691	0.9538	0.9698
XGBoost	60:40	0.9531	0.9390	0.9692	0.9539	0.9670
XGBoost	70:30	0.9530	0.9385	0.9694	0.9537	0.9701
XGBoost	80:20	0.9522	0.9378	0.9687	0.9530	0.9693
XGBoost	90:10	0.9522	0.9378	0.9687	0.9530	0.9695

Figure 4 visually encapsulates the stark difference between these two pipelines. The ROC curves for the CGAN-Ori pipeline (green and lime lines) are positioned far from the top-left corner, with low AUC scores. Their proximity to the random guess line provides a clear visual confirmation of the generalization failure. Conversely, the curves for the CGAN-CGAN pipeline (purple and fuchsia lines) are pushed strongly towards the optimal top-left corner, with outstanding AUC scores of approximately 0.97. This stark visual dichotomy is a central finding of this study, highlighting both the immense potential of CGANs in an ideal setting and the significant risk of overfitting when deploying these models against real-world, imbalanced data.

D. Comparison of Best Performing Models for Fraud Detection

This study evaluated the performance of RF and XGBoost classifiers across various training and testing dataset configurations, addressing class imbalance. The results, including accuracy for all configurations, are summarized in Table 10.

When both training and testing were done using the original imbalanced dataset, both RF and XGBoost classifiers showed high accuracy (97.37% and 97.38%, respectively), but their ability to detect fraudulent transactions was poor. While precision for fraud detection was moderate (80% for RF and 74% for XGBoost), the recall was alarmingly low (24%), indicating that a significant portion of fraudulent transactions were missed.

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 221–230
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46308

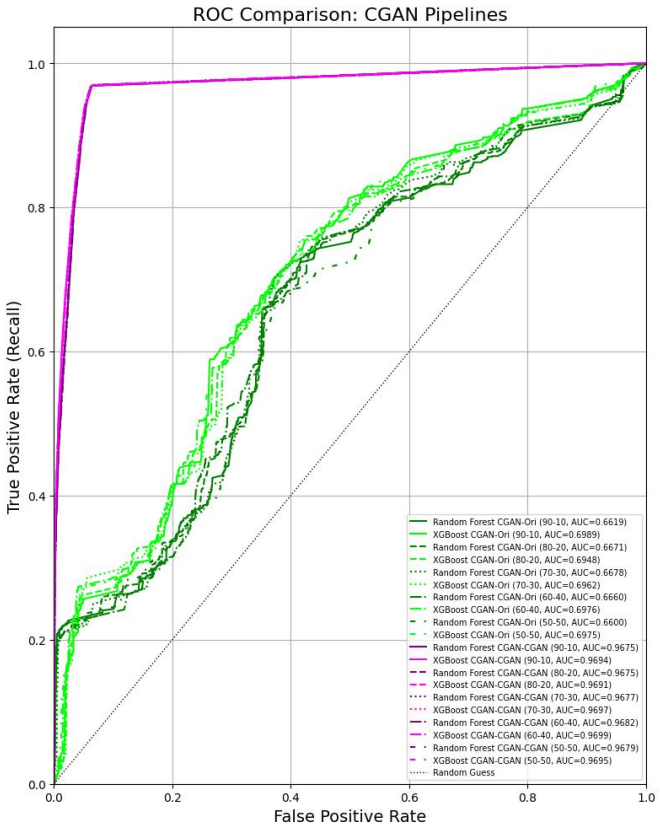

Fig. 4. ROC curves of CGAN-augmented dataset.

Table 10.
Comparative Analysis Across Different Training and Testing Configuration

Pipeline (Training -Testing)	Class.	Acc.	Prec.	Recall	F1-Score	ROC-AUC
Original-Original	RF	0,9737	0.8211	0.1939	0.3138	0.8999
	XGBoost	0.9738	0.7487	0.2324	0.3547	0.9025
SMOTE-Original	RF	0.8665	0.1597	0.7753	0.2648	0.8896
	XGBoost	0.8750	0.1692	0.7752	0.2777	0.8856
SMOTE-SMOTE	RF	0.8665	0.1597	0.7753	0.2648	0.8896
	XGBoost	0.8750	0.1692	0.7752	0.2777	0.8856
CGAN-Original	RF	0.9690	0.0000	0.0000	0.0000	0.5000
	XGBoost	0.9690	0.0000	0.0000	0.0000	0.4828
CGAN-CGAN	RF	0.9531	0.9389	0.9693	0.9539	0.9682
	XGBoost	0.9531	0.9390	0.9692	0.9539	0.9670

Class. = classifier, Acc. = accuracy, Prec. = precision

The F1-scores for both models were also low, 0.31 for RF and 0.35 for XGBoost, highlighting the challenge of fraud detection in imbalanced datasets, despite the high ROC-AUC scores (around 0.90). When the SMOTE-augmented dataset was used for training and tested on the original dataset, recall improved (78.87% for XGBoost), but precision remained very low (16.92%), and the F1-score stayed low at 0.27, indicating continued struggles with balancing precision and recall. However, the ROC-AUC scores improved slightly to 0.87 to 0.90.

Training and testing on the SMOTE-augmented dataset led to significant improvements, especially for XGBoost, with precision at 87.42%, recall at 82.24%, and an F1-score of 0.85, while the ROC-AUC score improved to 0.93, demonstrating strong fraud detection. RF also showed improvements, but XGBoost outperformed it. Training on the CGAN-augmented dataset achieved high accuracy (around 97%), but failed to detect fraudulent transactions on the original dataset, with zero precision, recall, and F1-scores. However, when both training and testing were done on the CGAN-augmented dataset, both models excelled, with XGBoost achieving precision (93.9%), recall (97%), F1-scores between 0.95 and 0.96, and a ROC-AUC score of 0.97.

The analysis shows that the best performance occurred when both training and testing were done with the CGAN-augmented dataset, especially for XGBoost, which achieved excellent precision, recall, F1-scores, and ROC-AUC. This setup demonstrated the model's effectiveness in detecting fraudulent transactions with augmented data, though performance on real-world data remained challenging, as seen with the CGAN-Original configuration. The SMOTE-augmented dataset improved recall but still struggled with low precision and F1-scores when tested on the original dataset. Therefore, a balanced dataset for both training and testing proved most effective for fraud detection.

IV. Discussion

A. *Interpretation of Results*

This study evaluates the performance of machine learning models for fraud detection by comparing the Original Dataset with SMOTE and CGAN-augmented datasets, where the original dataset's significant class imbalance leads to high accuracy (97.33% to 97.37%) for models like RF and XGBoost. However, accuracy alone is misleading in fraud detection, as it reflects the models' ability to predict non-fraudulent transactions rather than the rare fraudulent ones. Despite high accuracy, both classifiers had poor recall (18% to 24%), meaning many fraudulent transactions were missed. The precision was moderate for RF (around 80%) and lower for XGBoost (71% to 74%), with low F1-scores (0.31 for RF and 0.35 for XGBoost), revealing the imbalance between precision and recall. These results show that while accuracy is important, it’s insufficient for evaluating fraud detection models, as low recall and poor F1-scores highlight the need for metrics like precision and recall, with both RF and XGBoost struggling to identify fraudulent transactions.

The SMOTE technique balanced the dataset by generating synthetic fraudulent transactions, but fraud detection only

showed marginal improvements. For RF, accuracy remained high (86.29% to 86.66%), but precision stayed low (15.78% to 16.06%), and recall improved slightly (77.53% to 78.87%). XGBoost showed similar trends with slight improvements in accuracy (86.29% to 87.14%) and recall (78.87%), but precision remained low. These results highlight the challenges of fraud detection in imbalanced datasets, even with synthetic data augmentation.

The CGAN-augmented dataset addressed class imbalance by using a more sophisticated method, generating synthetic fraudulent transactions through a two-network system (generator and discriminator) for more realistic data. Unlike SMOTE, which interpolates between existing data points, CGAN's approach provides a more nuanced representation of real-world fraudulent transactions. However, when training was done using the CGAN-augmented dataset and tested on the original dataset, both RF and XGBoost classifiers faced significant issues. Despite achieving high accuracy (around 97%) for non-fraudulent transactions, the models had zero precision, recall, and F1-scores for fraudulent transactions across all test ratios, indicating they failed to detect any fraudulent transactions during testing. A comparison of real with synthetic feature distributions revealed that the CGAN generator overemphasized rare anomaly patterns and injected noise not present in genuine transactions. Classifiers hence overfit to these synthetic artifacts and could not generalize to authentic fraud cases. Incorporating distribution-alignment techniques (e.g., Maximum Mean Discrepancy regularization) or adding a diversity loss in GAN training could mitigate this mismatch.

This analysis reveals important insights into model performance. While both RF and XGBoost showed high accuracy with the original dataset, their ability to detect fraudulent transactions was hindered by class imbalance. The SMOTE-augmented dataset provided some improvements, but the models still struggled with fraud detection. Although the CGAN-augmented dataset showed promising results in training, it failed during testing on the original data, highlighting the need for further advancements in fraud detection, particularly in dealing with imbalanced datasets and synthetic data.

Real-time fraud detection systems are crucial in banking and e-commerce to mitigate financial losses from fraudulent activities. These systems leverage advanced machine learning and data analytics to enhance detection capabilities and improve operational efficiency. For example, [24] underscored the use of SMOTE alongside machine learning to boost accuracy in predicting financial fraud, while Mohammad et al. [25] demonstrate that integrating continuous analytics pipelines minimizes false positives and response latency. In practice, a SMOTE-augmented RF model can be deployed in a streaming pipeline (e.g., Kafka + Spark Streaming) to score incoming transactions with high recall within milliseconds. Conversely, a CGAN-augmented XGBoost model trained offline with GAN-based augmentation can run as a nightly batch job via RESTful microservices, delivering high-precision risk scores for investigator review.

B. Comparison with Previous Works

The findings of this study align with existing research on fraud detection in imbalanced datasets, while also highlighting the evolving effectiveness of augmentation methods and classifiers. SMOTE, a widely studied technique, has been shown to improve model performance by generating additional minority class samples through linear interpolation [3]. Our results support this, as SMOTE improved recall and F1-scores compared to the original dataset. However, SMOTE's performance was still outpaced by CGAN, especially in capturing the complex, nonlinear patterns of fraudulent transactions. This highlights the limitations of SMOTE's simplistic approach in the context of fraud detection.

CGANs, which generate synthetic data that mimics real-world distributions, have gained increasing attention for their ability to improve fraud detection. Recent studies emphasize CGAN's success due to its adversarial training framework, which produces high-fidelity minority class samples [5]. Our results further validate these claims, with CGAN-augmented datasets consistently outperforming SMOTE and the original dataset in terms of recall, precision, and F1-score. Notably, XGBoost paired with CGAN achieved a recall of 96.88% and an F1-score of 95.31%, surpassing state-of-the-art results in studies employing GAN-based augmentation techniques [26].

The superior performance of CGAN compared to SMOTE aligns with recent trends in the field, highlighting the importance of data quality over quantity. Unlike SMOTE, which uses interpolation and may fail to capture the complexities of fraud, CGAN generates synthetic data through iterative adversarial learning, producing more nuanced samples of fraudulent behavior [7]. This is especially beneficial for classifiers like XGBoost and RF, which are adept at extracting complex patterns from structured data [5]. The combination of CGAN with these advanced classifiers demonstrates the transformative potential of using sophisticated data augmentation methods alongside robust machine learning algorithms.

This study challenges previous research by critically evaluating models trained on the original dataset, showing that high accuracy on imbalanced datasets can be misleading due to low recall values [27]. For instance, RF and XGBoost achieved high accuracy but had recall rates of only 19.39% and 24.54%, respectively, highlighting the importance of focusing on recall and other minority-class metrics in fraud detection, a viewpoint increasingly emphasized in the literature [8], [12]. Focusing solely on accuracy can obscure the model's failure to identify important instances, highlighting the necessity for more comprehensive metrics like precision, recall, F1-score, and ROC-AUC in assessing fraud detection systems. The challenge of data imbalance and its impact on model evaluation has been widely debated, with numerous studies recommending alternative evaluation metrics [28].

The results of this study demonstrate the effectiveness of augmentation techniques in addressing class imbalance, with CGAN consistently outperforming both SMOTE and the

Applied Information System and Management (AISM)
Volume 8, (2) 2025, p. 221–230
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i2.46308

original dataset across all metrics for RF and XGBoost classifiers. CGAN's generative approach, which produces diverse and realistic synthetic samples, enabled the classifiers to better detect minority class instances. While SMOTE showed some improvements over the original dataset, it was less effective, particularly in recall and F1-score. The original imbalanced dataset consistently underperformed, emphasizing the need for balancing techniques in fraud detection.

For RF, CGAN was the most effective augmentation method, achieving the highest recall and F1-score, enhancing the model's ability to learn from underrepresented samples by generating synthetic data similar to the minority class, while SMOTE improved recall and precision but produced less varied samples, limiting its effectiveness. Similarly, for XGBoost, CGAN outperformed SMOTE across all metrics, particularly in ROC-AUC and F1-score. XGBoost's advanced boosting mechanism likely benefited from CGAN's high-quality synthetic data, improving minority class detection. These findings highlight the limitations of traditional evaluation metrics and underscore the importance of techniques like CGAN in imbalanced data applications, in line with previous studies that emphasize the role of specialized strategies in such cases [29].

While CGAN generated samples yield higher precision and F1-scores on matched data, they come at the cost of substantially longer training time, greater implementation complexity, and a heightened risk of overfitting to synthetic artifacts. In contrast, SMOTE offers a lightweight, deterministic oversampling approach that integrates seamlessly into real-time pipelines but produces less diverse fraud examples, limiting its peak recall improvements.

To our knowledge, this is the first study to perform a head-to-head comparison of SMOTE with CGAN augmentation under identical RF and XGBoost settings on a large, real-world e-commerce fraud dataset. We also provide the inaugural root-cause analysis of CGAN's generalization failure identifying key distributional mismatches and deliver a fully reproducible evaluation pipeline with paired t-tests and confidence intervals for robust statistical benchmarking.

C. *Limitations*

Despite promising results, this study has limitations, mainly due to the high computational demands of techniques like CGAN. While CGAN improved sample generation for the minority class, its training process required substantial computational power and time, especially on hardware like the MacBook Air M1, which limited scalability and efficiency and hindered more complex hyperparameter tuning and iterative refinements [5]. To address this, future work should leverage distributed training frameworks on cloud platforms (e.g., AWS SageMaker, Azure ML) or multi-GPU servers to accelerate GAN convergence and enable more comprehensive tuning. Moreover, e-commerce dataset while extensive, may not reflect fraud dynamics in other domains such as healthcare or cybersecurity, where attack patterns differ. Expanding evaluations to cross-industry datasets and employing federated learning setups can improve generalizability and privacy compliance. Finally, both SMOTE and CGAN risk overfitting to synthetic artifacts, potentially missing novel fraud strategies. Incorporating domain-adaptation losses, adversarial validation, or online continual learning pipelines will help align synthetic and real distributions and adapt models to emerging fraud patterns. These enhancements will strengthen the robustness and applicability of augmentation-based fraud detection in real-world settings.

V. CONCLUSION

This study provided valuable insights into the relationship between data augmentation techniques and classifier algorithms for fraud detection on imbalanced datasets. While CGAN consistently delivered superior precision, recall, and F1-scores demonstrating its ability to generate high-fidelity minority samples that enhance RF and XGBoost performance SMOTE still offers meaningful gains over the original data by rapidly improving recall through deterministic oversampling. For practitioners, the recommendation to deploy SMOTE-augmented models in real-time or low-latency environments (such as streaming payments and API-based checks), where its CPU-only implementation maximizes fraud capture within minutes, and reserving CGAN-augmented models for batch or offline scoring (for example, nightly risk assessments), where its richer synthetic diversity and higher precision justify GPU training overhead. A hybrid architecture SMOTE for immediate alerts and CGAN for in-depth analysis can balance false-positive and false-negative trade-offs according to operational priorities. By clarifying these use cases, our work moves beyond theoretical comparison to deliver concrete, context-aware guidance for data scientists and fintech engineers. Future research may extend our framework by exploring variational autoencoders for augmentation, leveraging CNNs or RNNs to capture temporal fraud patterns, and scaling evaluations across sectors such as healthcare and cybersecurity. Integrating adaptive, real-time learning mechanisms and explainable AI techniques will further bolster system robustness and stakeholder trust in deployed fraud detection solutions.

REFERENCES

[1] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, no. July, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[2] T. R. Noviandy, G. M. Idroes, A. Maulana, I. Hardi, E. S. Ringga, and R. Idroes, "Credit card fraud detection for contemporary financial management using XGBoost-driven machine learning and data augmentation techniques," *Indatu J. Manag. Account.*, vol. 1, no. 1, pp. 29–35, 2023, doi: 10.60084/ijma.v1i1.78.

[3] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based oversampling

data-point approach to solving the credit card data imbalance problem in financial fraud detection," *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 277–286, 2021, doi: 10.12785/ijcds/100128.

[4] D. P. Kadam, R. S. Chiparikar, M. A. Kamble, and M. H. Attarde, "Machine learning approaches to credit card fraud detection," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 4, pp. 2802–2807, 2024, doi: 10.22214/ijraset.2024.60531.

[5] M. J. Rahman and H. Zhu, "Detecting accounting fraud in family firms: Evidence from machine learning approaches," *Adv. Account.*, vol. 64, no. March, Art. no. 100722, 2024, doi: 10.1016/j.adiac.2023.100722.

[6] M. Zhu, Y. Zhang, Y. Gong, C. Xu, and Y. Xiang, "Enhancing credit card fraud detection: A neural network and SMOTE integrated approach," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 02, pp. 23–30, 2024, doi: 10.53469/jtpes.2024.04(02).04.

[7] E. Wu, H. Cui, and R. E. Welsch, "Dual autoencoders generative adversarial network for imbalanced classification problem," *IEEE Access*, vol. 8, pp. 91265–91275, 2020, doi: 10.1109/ACCESS.2020.2994327.

[8] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *J. Big Data*, vol. 10, no. 1, pp. 1-17, 2023, doi: 10.1186/s40537-023-00684-w.

[9] I. de Zarzà, J. de Curtò, and C. T. Calafate, "Optimizing neural networks for imbalanced data," *Electron.*, vol. 12, no. 12, pp. 1–26, 2023, doi: 10.3390/electronics12122674.

[10] D. P. Prabha and C. V Priscilla, "Probabilistic XGBoost threshold classification with autoencoder for credit card fraud detection," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 8s, pp. 528–537, 2023, doi: 10.17762/ijritcc.v11i8s.7234.

[11] R. M. Mathew and R. Gunasundari, "A hybrid resampling approach for multiclass skewed datasets and experimental analysis with diverse classifier models," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 10, pp. 1108–1114, 2023, doi: 10.17762/ijritcc.v11i10.8631.

[12] E. Pan, "Machine learning in financial transaction fraud detection and prevention," *Trans. Econ. Bus. Manag. Res.*, vol. 5, pp. 243–249, 2024, doi: 10.62051/16r3aa10.

[13] S. F. Farabi, M. Prabha, M. Alam, Z. Hossan, Md. Arif, R. Islam, A. Uddin, M. Bhuiyan, Md. Zinnat, and A. Biswas, "Enhancing credit card fraud detection: A comprehensive study of machine learning algorithms and performance evaluation," *J. Bus. Manag. Stud.*, vol. 6, no. 3, pp. 252–259, 2024, doi: 10.32996/jbms.2024.6.13.21.

[14] G. Airlangga, "Evaluating the efficacy of machine learning models in Credit Card Fraud Detection," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 2, pp. 829–837, 2024, doi: 10.47709/cnahpc.v6i2.3814.

[15] D. Lin, "An empirical analysis of machine learning for fraud detection in diverse financial scenarios," *Adv. Econ. Manag. Polit. Sci.*, vol. 42, no. 1, pp. 202–216, 2023, doi: 10.54254/2754-1169/42/20232110.

[16] A. K. Nandi, K. K. Randhawa, H. S. Chua, M. Seera, and C. P. Lim, "Credit card fraud detection using a hierarchical behavior-knowledge space model," *PLoS One*, vol. 17, no. 1, Art. no. e0260579, 2022, doi: 10.1371/journal.pone.0260579.

[17] P. Li, "Credit card fraud detection based on random forest model," *Acad. J. Comput. Inf. Sci.*, vol. 5, no. 13, pp. 55–61, 2022, doi: 10.25236/ajcis.2022.051309.

[18] 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, no. April, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[19] H. Estelami and K. Liu, "Content analysis of american consumers' credit card fraud complaints filed with the consumer financial protection bureau," *J. Financ. Crime*, vol. 31, no. 3, pp. 618–628, 2023, doi: 10.1108/jfc-03-2023-0070.

[20] M. Zhang, "Identification and analysis of credit card fraud based on machine learning methods," *Adv. Econ. Manag. Polit. Sci.*, vol. 94, no. 1, pp. 109–122, 2024, doi: 10.54254/2754-1169/94/2024ox0182.

[21] M. S. Khatun, B. R. Alam, M. Taslim, and M. A. Hossain, "Handling class imbalance in credit card fraud using various sampling techniques," *Am. J. Multidiscip. Res. Innov.*, vol. 1, no. 4, pp. 160–168, 2022, doi: 10.54536/ajmri.v1i4.633.

[22] S. Bhardwaj and S. Gupta, "Effects of feature selection with machine learning algorithms in detection of credit card fraud," *Int. J. Eng. Res. Comput. Sci. Eng.*, vol. 9, no. 7, pp. 46–51, 2022, doi: 10.36647/ijercse/09.07.art011.

[23] Q. Zeng, L. Lin, R. Jiang, W. Huang, and D. Lin, "NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks," *Inf. Process. Manag.*, vol. 62, no. 1, pp. 1-15, 2025, doi: 10.1016/j.ipm.2024.103916.

[24] Z. Zhao and T. Bai, "Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms," *Entropy*, vol. 24, no. 8, pp. 1-17, 2022, doi: 10.3390/e24081157.

[25] N. Mohammad, M. A. U. Imran, M. Prabha, S. Sharmin, and R. Khatoon, "Combating banking fraud with it: integrating machine learning and data analytics," *Am. J. Manag. Econ. Innov.*, vol. 6, no. 7, pp. 39–56, 2024, doi: 10.37547/tajmei/volume06issue07-04.

[26] P. Verma and P. Tyagi, "Analysis of supervised machine learning algorithms in the context of fraud detection," *Ecs Trans.*, vol. 107, no. 1, pp. 7189–7200, 2022, doi: 10.1149/10701.7189ecst.

[27] Y. Bing Chu, Z. Min Lim, B. Keane, P. Hao Kong, A. Rafat Elkilany, and O. Hisham Abusetta, "Credit card fraud detection on original european credit card holder dataset using ensemble machine learning technique," *J. Cyber Secur.*, vol. 5, no. November, pp. 33–46, 2023, doi: 10.32604/jcs.2023.045422.

[28] Haibo He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[29] S. F. Pratama, "Fraudulent transaction detection in online systems using random forest and gradient boosting," *J. Cyber Law*, vol. 1, no. 1, pp. 88–115, 2025, doi: 10.63913/jcl.v1i1.5.