FRAUD DETECTION FROM BANK CARD FINANCIAL TRANSACTIONS IN NIGERIA: A COMPARATIVE EVALUATION OF CLASS IMBALANCE HANDLING AND MACHINE LEARNING MODELS

SULAIMON ADEBAYO BASHIR[1], OLADAYO TOSIN AKINWANDE[2], SOLOMOM ADELOWO ADEPOJU[3], ADERIIKE OPEYEMI ABISOYE [4] & LAWAL O. LAWAL[5]
Department of Computer Science, Federal University of Technology Minna[1,3,4,5]
Department of Software Engineering, Veritas University Abuja[2]
Email: bashirsulaimon[1] , o.abisoye[3], solo.adepoju[4],lawal.lawa[5]]@futminna.edu.ng, akinwandeo@veritas.edu.ng[2]

Abstract
This study investigates the performance of five machine learning models: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) for fraud detection in real-world Nigerian bank card transaction data. The study addresses the class imbalance problem through two data balancing strategies- Synthetic Minority Oversampling Technique (SMOTE) and random downsampling. The results show that for the imbalanced dataset logistics regression achieved highest average accuracy of 99.99% followed by Random Forest , XGBoost, SVM and KNN with 99.98, 99.92, 98.69 and 98.54% respectively. While for the SMOTE-balanced dataset Random Forest achieved highest average accuracy of 99.99% followed by Logistic Regression, XGBoost, KNN and SVM with 99.98, 99.92, 98.80 and 97.92% respectively. The evaluation on downsampled dataset revealed that Logistic Regression, Random Forest, XGBoost , KNN and SVM attained average accuracies of 99.98, 99.93, 99.86, 96.90 and 98.05. These findings imply that the five machine learning models performed well on all the 3 variations of the dataset. This can be attributed to the quality and size of the dataset that makes the imbalance not to have negative effects on the performance of the models.

Keywords: Fraud Detection, Bank Cards, Financial Transactions, Class Imbalance, SMOTE, Machine Learning, Random Forest, XGBoost, KNN

Introduction

The adoption of electronic banking in Nigeria has facilitated financial inclusion but also introduced significant risks associated with card-based financial frauds. As financial transactions increasingly move online, fraudsters exploit vulnerabilities in digital infrastructures, making fraud detection a critical challenge for financial institutions (Mosa *et al.,* 2024). Nigeria's socio-economic environment also plays a crucial role in the prevalence of credit card fraud. The rapid adoption of online marketing and e-commerce has led to an increase in fraudulent activities, with millions of Nigerians restricted from using local debit and credit cards internationally due to fraud concerns (Njoku *et al,* 2024). The łack of robust regulatory frameworks and the high cost of implementing advanced fraud detection systems further exacerbate the problem (Misol & Agbadua, 2022).

Credit card fraud is a general phrase that refers to a situation in which unauthorised users obtain an individual's credit card information in order to make purchases, conduct other transactions, or transfer funds to another location (Shah & Makwana, 2023). The Central Bank of Nigeria has implemented measures such as the Bank Verification Number (BVN) to trace and freeze accounts linked to fraudulent activities (CBN, 2018; Adewale, 2020). However, the effectiveness of these measures is often hindered by the sophistication of fraudsters and the limited awareness among cardholders about fraud prevention (Misol & Agbadua, 2022). The wide acceptance of cards as preferred mode of payment across a variety of transaction platforms have also, attracted scammers with a great rise in the number of threats, successful

attacks and fraudulent activities (Otorokpo *et al.,* 2024). It is now imperative and critical for financial institutions to advance efforts to enhance their fraud detection and prevention systems aimed at mitigating further losses to fraudsters, who target their systems and customers for fraudulent financial gains. However, fraudulent activities in bank card transactions are often rare events compared to legitimate transactions, resulting in highly imbalanced datasets that hinder performance of machine learning models (Dal Pozzolo *et al.,* 2014). It has also been noted that data imbalance issues can potentially impede the effectiveness of the learning process for fraud detection (Feng & Kim, 2024) which has made the research of class imbalance dataset an important one.

This paper investigated the effect of class imbalance on the performance of selected machine learning models on the detection of fraud in bank card transactions data collected from a financial institution in Nigeria. The rest of the paper is organized as follows: Section 2 reviewed the related literature. Section 3 described the experiment. Section 4 presented the result. Conclusion was also presented in section 6. Section 6 outlines future direction of this work.

Related Work

Prior research has addressed fraud detection using various supervised and unsupervised machine learning algorithms, including logistic regression, decision trees, support vector machines, and neural networks. However, class imbalance remains a core obstacle. Techniques such as SMOTE (Synthetic Minority Oversampling Technique) (*Chawla et al.,* 2002), random Undersampling (Dal Pozzolo *et al.,* 2015), and ensemble methods have been proposed to improve classifier sensitivity toward minority (fraudulent) classes (Carcillo *et al.,* 2019).

Traditional machine learning approaches have been widely used for credit card fraud detection in Nigeria. These methods include supervised and unsupervised learning techniques, each with its strengths and weaknesses. Studies focused on the Nigerian financial context are limited, even though the country presents unique challenges such as high rates of cashless transactions, rapid growth in mobile banking, and diverse banking customer demographics. This study contributes to closing this gap by applying comparative imbalance handling techniques specifically to Nigerian bank card transactions. Supervised learning algorithms are the most commonly used methods for fraud detection. These algorithms learn from labelled data, where each transaction is marked as either fraudulent or legitimate. Common supervised learning techniques include: Random Forest (RF) has been shown to perform well in fraud detection tasks, especially when combined with data balancing techniques like SMOTE. Random Forest has achieved an accuracy of 0.9802 when used with SMOTE (Otorokpo *et al.,* 2024). Support Vector Machines (SVM) are effective in high-dimensional data and have been used to detect fraudulent transactions with good accuracy (Khedkar and Gupta, 2024; Jayabalan and Shiksha, 2024). Logistic Regression (LR) is a simple yet effective method for fraud detection. It has been used in combination with various sampling techniques to improve performance (Jayabalan & Shiksha, 2024; Jain *et al*., 2024). Other supervised learning technique which has been used in fraud detection are decision Trees and K-Nearest Neighbors (KNN) which are easy to interpret and have been used in fraud detection systems also. However, they often require ensemble methods to improve their performance (Jayabalan and Shiksha, 2024; Farabi *et al.,* 2024). Unlike decision trees, K-Nearest Neighbors (KNN) performance is often lower compared to other algorithms like RF and SVM (Farabi *et al.,* 2024; Jain *et al.,* 2024).

Despite the success of some traditional machine learning models in detecting card fraud detection, the problem of class imbalance still persists. Class imbalance is a significant challenge in credit card fraud detection (Dal Pozzolo *et al.,* 2018), as fraudulent transactions are often a small minority compared to legitimate transactions. Various techniques have been

proposed to address this issue. Techniques such as Data-level techniques, algorithmic techniques and hybrid approaches. Data-level techniques aim to balance the dataset before training the model and common methods include Synthetic Minority Oversampling Technique (SMOTE) that generate synthetic samples of the minority class and Random Undersampling (RUS) which reduces the number of majority class samples to balance the dataset. It is often used in combination with oversampling techniques (Berkins & Karthick 2022; Penumala, 2024).

Algorithmic techniques modify the learning process to minimize bias towards the majority class. Cost-sensitive learning and ensemble methods are popular algorithmic techniques used in class imbalance problem. Cost-sensitive learning assigns higher costs to misclassifying minority class samples. It has been used in fraud detection to improve the detection of fraudulent transactions (Priatna *et al.,* 2024) while Ensemble methods combine multiple models to improve overall performance (Otorokpo *et al.,* 2024; Manda *et al.,* 2024). Ensemble methods like Random Forest and XGBoost are effective in handling class imbalance. Hybrid approaches combine data-level and algorithmic techniques to create robust models. Stacking ensemble models and deep learning techniques have been used (Lebichot *et al.,* 2019). Stacking ensemble models with SMOTE-ENN have been used to improve fraud detection performance. These models have achieved high accuracy and AUC scores in imbalanced datasets (Jyoti *et al.,* 2024). Deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used in combination with SMOTE to improve feature extraction and classification performance (Elmangoush *et al.,* 2024).

Method

Dataset Description and Exploratory Analysis

The dataset consists of anonymised bank card financial transactions from a Nigerian banking institution. The data was obtained through a collaborative research with a team of data analysts from the bank. Each record represents a card-based transaction with features related to the customer profile, card characteristics, and transaction behaviour. Features include numeric and categorical variables such as Transaction Amount, Customer Age, Cards, CardType, TransactionType, Domain, and Outcome. The target feature Outcome indicates whether a transaction is fraudulent (0) or legitimate (1). The detailed description of the 16 features in the dataset is presented in Table 1.

Table 1: Dataset Attributes

No.	Features	Description	Data Type
1	CustomerAge	Customer age	Numerical
2	ATM	Automated Teller Machine (ATM) transaction limit	Numerical
3	POSWEBLimit	Point of Sale(POS)/Web transaction limit	Numerical
4	CreditLimit	Credit card limit	Numerical
5	Amount	Transaction amount	Numerical
6	AverageIncomeExpendicture	Average income/expenditure	Numerical
7	NewBalance	Post-transaction balance	Numerical
8	OldBalance	Pre-transaction balance	Numerical
9	Gender	Customer gender	Categorical
10	Marital Status	Customer marital status	Categorical
11	Cards	Card type	Categorical
12	CardColour	Card color	Categorical
13	CardType	Card brand/type	Categorical

14	TransactionType	Type of transaction	Categorical
15	Domain	Transaction domain	Categorical
16	Outcome	Label / Target	Binary (0 or 1)

Exploratory Data Analysis (EDA) of the Bank Card Transaction Dataset

Further exploratory analysis was performed on the dataset to identify pertinent information about the data values in the dataset. The attributes were categorised into two: numeric and categorical attributes. The EDA of the two are presented in Table 2 and Table 3 respectively. The exploratory analysis of the numeric part of the dataset has 8 features as shown in Table 2. All the values are present in the data except for *CustomerAge* feature which has 8,851 (about 23.85%) missing values. These missing values are addressed in the preprocessing step as discussed in Section 3.3. The values in the CustomerAge values range from 18 to 85 years, with a mean age of 39.2 years (SD = 20.1). Most customers (50%) are below 29 years, showing that younger customers dominate the dataset.

Table 2: Numeric Data Exploratory Analysis

Feature	Missing values	count	min	Max	Mean	std
Customer Age	8851	28246	18	85	39.2	20.1
ATMLimit	0	37097	120,000	150,000	140,870.7	13,803.6
POSWEB Limit	0	37097	1,200,000	4,000,000	2,452,101.2	1,066,813.9
Credit Limit	0	37097	150,000	600,000	335,101.2	169,488.7
Amount	0	37097	100,003	999,956	550,261.0	260,629.8
Average Income Expenditure	0	37097	100,017	399,971	227,387.0	78,977.1
New Balance	0	37097	-897,378	1,591,355	294,806.6	623,961.4
OldBalance	0	37097	100,001	599,990	350,202.8	144,781.9

ATM, POSWEBLimit, and CreditLimit represent transaction limits. *ATM* limits for different customers are mostly close to the maximum of 150,000 since 75 percentile of the ATM Limits values are less than 150,000, with very little variation (SD = 13,803.63). *POSWEBLimit* on the other hand, shows much wider differences between customers (from 1.2M to 4M; SD = 1,066,813.92) while C*reditLimit* averages 335,101, with high variation (SD = 169,488.72), likely reflecting differences in customer credit levels.

In the same vein, Amount (transaction amount) ranges from 100,003 to 999,956, with a mean of 550,261 (SD = 260,629.81), showing large differences in transaction sizes while Average Income Expendicture shows a mean income expenditure of 227,386 (SD = 78,977), with less spread compared to transaction amounts. The New Balance and Old Balance features represent balances after and before transactions. *New Balance* shows wide variation (from 897,378 to 1,591,355), including negative balances which may indicate overdrafts or debts as these values cut across different cards (debit, credit and prepaid). *Old Balance* ranges from 100,001 to 599,990, and is generally more stable (SD = 144,781.95).

Overall, these numeric features show both stable patterns and wide variations. The large spread in transaction values and balances is important for fraud detection because unusual or extreme values may signal fraudulent activities. Because transaction amounts carry meaning in fraud detection, scaling or normalization was not applied to these numeric features to keep their original values for the models.

Table 3: Categorical Features Data Exploratory Analysis

Feature	Missing	Unique Categories	Most Common	Most Common Count	Distribution
Gender	0	2	Male	23186	{'Male': 23186, 'Female': 13911}
Marital Status	0	4	Married	17257	{'Married': 17257, 'Single': 14393, 'Divorced': 2743, 'Unknown': 2704}
Cards	0	3	Debit	18550	{'Debit': 18550, 'Credit': 11289, 'Prepaid': 7258}
CardColour	0	2	Gold	18550	{'Gold': 18550, 'White': 18547}
CardType	0	3	Verve	18550	{'Verve': 18550, 'Visa': 11289, 'MasterCard': 7258}
Transaction Type	0	2	Debit	20440	{'Debit': 20440, 'Credit': 16657}
Domain	0	2	International	26497	{'International': 26497, 'Local': 10600}
Outcome	0	2	1	27370	{1: 27370, 0: 9727}

Eight categorical features related to customer demographics, card properties, transaction types, and fraud outcomes are identified in the dataset as shown in Table 3. As indicated in the table, there are no missing values in any of the categorical features. Among these, *Gender* is divided into Male (62.5%) and Female (37.5%), indicating male customers are more prevalent. For *Marital Status*, Married customers formed the largest group (46.5%), followed by Single (38.8%), while Divorced and Unknown categories appear less frequently. Regarding card usage, *Cards* shows that Debit cards are most common (50%), while Credit (30.4%) and Prepaid (19.6%) cards are less represented. *CardColour* is evenly split between Gold (50%) and White (50%), whereas *CardType* reveals Verve as the dominant card type (50%), ahead of Visa (30.4%) and MasterCard (19.6%). In terms of transaction methods, *TransactionType* indicates a slight dominance of Debit transactions (55%) over Credit transactions (45%). The *Domain* feature shows that most transactions are International (71.4%), with Local transactions accounting for 28.6%. Finally, the target variable *Outcome* is highly imbalanced, with 73.8% of transactions labeled as legitimate (Outcome = 1) and only 26.2% marked as fraudulent (Outcome = 0), highlighting a significant class imbalance that was carefully addressed during preprocessing. The imbalance in the target variable indicates that actual fraud cases are less prevalent compared to legitimate transactions reflecting what is usually obtainable in many real-life anomaly scenarios.

Data Pre-processing

With the exploratory data analysis revealing the characteristics and limitations of the dataset that include missing values, imbalance target feature and the need to encode textual

categorical features to numeric encoding; this section presents the pre-processing performed on the dataset to make it more amenable to machine learning algorithms model induction. To address the missing values in the CustomerAge feature, we applied median imputation technique for handling missing data. The median imputation process involves replacing missing values in a feature with the median of the observed (non-missing) values of that same feature. The median of a given set of data point $X = \{x_1, x_2, \ldots x_n\}$ is computed according to Equation 1.

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & if\ n\ is\ odd \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & if\ n\ is\ even \end{cases} \quad (1)$$

Categorical features were label encoded into integers to enable machine learning models to process them. Each categorical value is assigned value ranging from 0 to number of distinct categories. For each categorical feature V with unique values $\{v_1, v_1, \ldots, v_k\}$, a mapping function $f: C \rightarrow \mathrm{Z}$ was created such that:

$$f(v_i) = i - 1 \quad \forall \quad i = 1, 2, \ldots, k \quad (2)$$

To address the class imbalance of the dataset whereby samples with outcomes labelled fraud transactions (0) are 9,727 and legitimate transactions (1) are 27,370 representing 26.2% and 73.8% respectively, two data balancing techniques were applied:

I. **SMOTE**: Synthetic samples for the minority class were generated using nearest neighbor interpolation.
II. **Random Downsampling**: Randomly reduces the majority class to the same size as the minority class.

At the end of the pre-processing, three pre-processed datasets were prepared: Unbalanced, SMOTE-balanced, and Downsampled-balanced. The distributions of the number of samples for normal(1) and fraud (0) for the three datasets are shown in Figure 1.

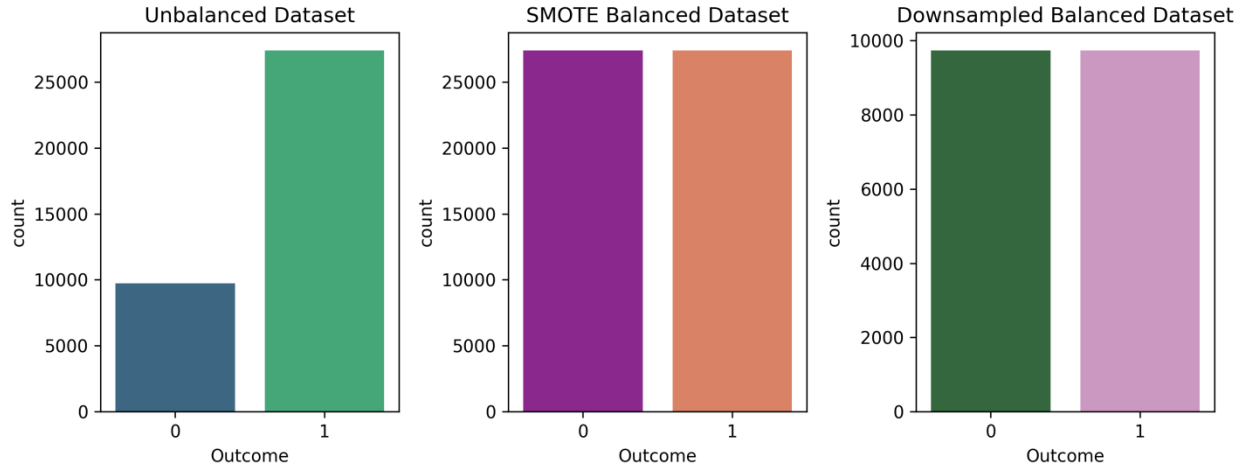

Figure 1: Three Pre-processed Datasets

Experimental Setup

Five machine learning models were trained and evaluated. They include Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). These algorithms were selected because they have been reported extensively to perform well for numeric data (Bashir *et al.,* 2015; Bashir *et al.,* 2016; Umar *et al.,* 2019).

Furthermore, the experiment setup involves training each model with each of the pre-processed dataset that included i) the original imbalanced dataset with customer age missing

value imputed via median imputation method ii) balanced dataset using SMOTE and iii) downsampled dataset that eliminate records or samples with missing values. Each of these datasets were partitioned into 70% training and 30% test samples to evaluate the trained model. Of course, KNN does not create a model as it is instance-based learning. Nonetheless, the same data partition was used for KNN model as well.

Evaluation Metrics

The evaluation metrics employed to quantify the performance of the models across the different datasets are described as follows:

Let:

C_1: Positive class (label 1)
C_0: Negative class (label 0)
TP: True Positives (samples correctly predicted as C_1)
TN: True Negatives (samples correctly predicted as C_0)
FP: False Positives (samples incorrectly predicted as C_1 while beingC_0)
FN: False Negatives (samples incorrectly predicted as C0 while beingC_1)

The following performance metrics were computed:

Accuracy: The overall proportion of correctly classified instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Precision (for class C_1): The proportion of predicted C_1 instances that are correctly classified.

$$Precision_{C_1} = \frac{TP}{TP + FP} \quad (4)$$

Precision (for class C_0): The proportion of predicted C_0 instances that are correctly classified.

$$Precision_{C_0} = \frac{TN}{TN + FN} \quad (5)$$

Recall (Sensitivity, True Positive Rate (TPR), for class C_1): The proportion of actual C_1 instances correctly predicted.

$$Recall_{C_1} = TPR = \frac{TP}{TP + FN} \quad (6)$$

Recall (Specificity, True Negative Rate (TNR), for class C_0): The proportion of actual C_0 instances correctly predicted.

$$Recall_{C_0} = TNR = \frac{TN}{TN + FP} \quad (7)$$

F1-Score (for class C_1**):** The harmonic mean of precision and recall for class C_1:

$$F1_{C_1} = 2 \times \frac{Precision_{C_1} \times Recall_{C_1}}{Precision_{C_1} + Recall_{C_1}} \quad (8)$$

F1-Score (for class C_0 **):** Similarly computed for class C_0:

$$F1_{C_0} = 2 \times \frac{Precision_{C_0} \times Recall_{C_0}}{Precision_{C_0} + Recall_{C_{01}}} \quad (9)$$

False Positive Rate (FPR, for class C_0)

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

False Negative Rate (FNR, for class C_1)

$$FNR = \frac{FN}{FN + TP} \quad (11)$$

Area Under the ROC Curve (AUC-ROC): The AUC quantifies the overall ability of the classifier to distinguish between the positive and negative classes across all possible classification thresholds. It is computed as the area under the curve plotting TPR vs FPR.

Results and Discussions

The performance evaluation of the five classifiers across the three datasets namely: imbalanced, SMOTE-balanced and down sampled reveals several notable patterns. Table 4 shows the performance evaluation of the fraudulent transactions (Class 0) data where the Random Forest and XGBoost consistently achieved very high precision and recall across all data scenarios as depicted in Figure 2 and Figure 3 respectively. Logistic Regression also performed well, closely matching the ensemble models in most cases. On the unbalanced dataset, Logistic Regression, Random Forest, and XGBoost all maintained precision and recall above 99.8%, while SVM and KNN exhibited substantially lower performance, with SVM yielding an F1-score of 97.45% and KNN 97.16%, indicating their limited capacity to handle severe class imbalance without explicit balancing.

The application of SMOTE substantially improved minority class performance for SVM and KNN. With the increase in fraud sample support after oversampling, SVM achieved perfect recall (1.000) and KNN reached an F1-score of 98.81%. This underscores SMOTE's effectiveness in enriching the minority class representation and enabling these algorithms to better learn fraud patterns. The ensemble models remained stable under SMOTE, continuing to produce near-perfect scores, reaffirming their robustness to both original and artificially balanced datasets. Similarly, downsampling the majority class improved minority class performance for SVM and KNN relative to the unbalanced dataset, although their results still lagged behind the ensembles. Notably, SVM achieved a recall of 100% but with lower precision, resulting in an F1-score of 96.90%, while KNN reached an F1-score of 98.01% on the downsampled data.

Table 4: Performance Evaluation Result on Class 0 (Fraud) Prediction

Dataset	Model	Class 0 Precision	Class 0 Recall	Class 0 F1 Score	Class 0 Support
Imbalanced	Logistic Regression	0.9997	1.0000	0.9998	2884
	Random Forest	0.9997	0.9997	0.9997	2884
	XGBoost	0.9983	0.9986	0.9984	2884
	SVM	0.9810	0.9681	0.9745	2884
	KNN	0.9769	0.9664	0.9716	2884
SMOTE-balanced	Logistic Regression	0.9999	0.9998	0.9998	8170
	Random Forest	0.9999	0.9999	0.9999	8170
	XGBoost	0.9994	0.9990	0.9992	8170
	SVM	0.9599	1.0000	0.9796	8170
	KNN	0.9791	0.9972	0.9881	8170
Downsampled	Logistic Regression	0.9996	1.0000	0.9998	2829
	Random Forest	0.9986	1.0000	0.9993	2829
	XGBoost	0.9982	0.9989	0.9986	2829
	SVM	0.9399	1.0000	0.9690	2829
	KNN	0.9667	0.9940	0.9801	2829

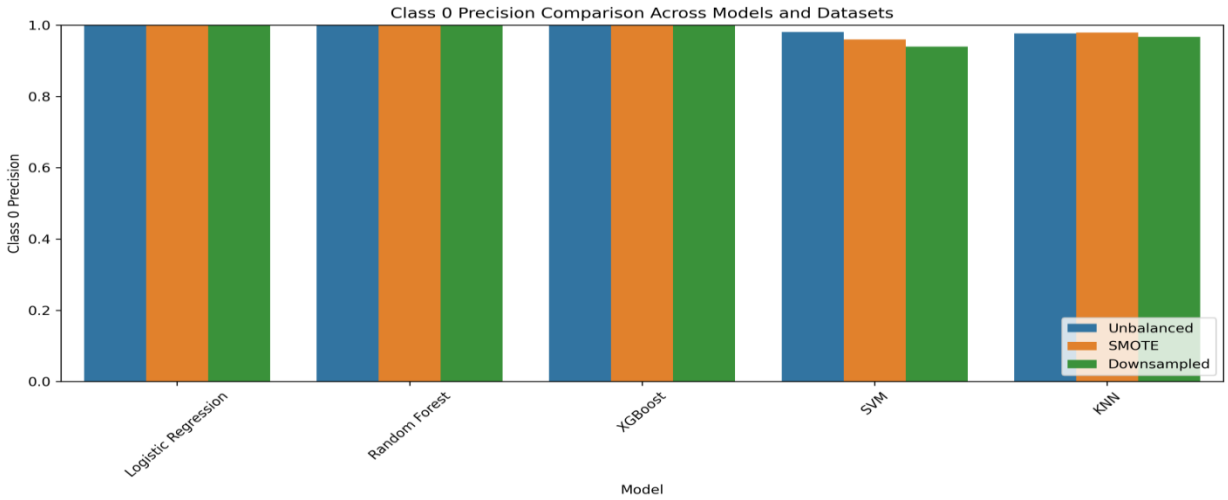

Figure 2: Precision Metric Performance of the Models on Fraudulent Transaction Data

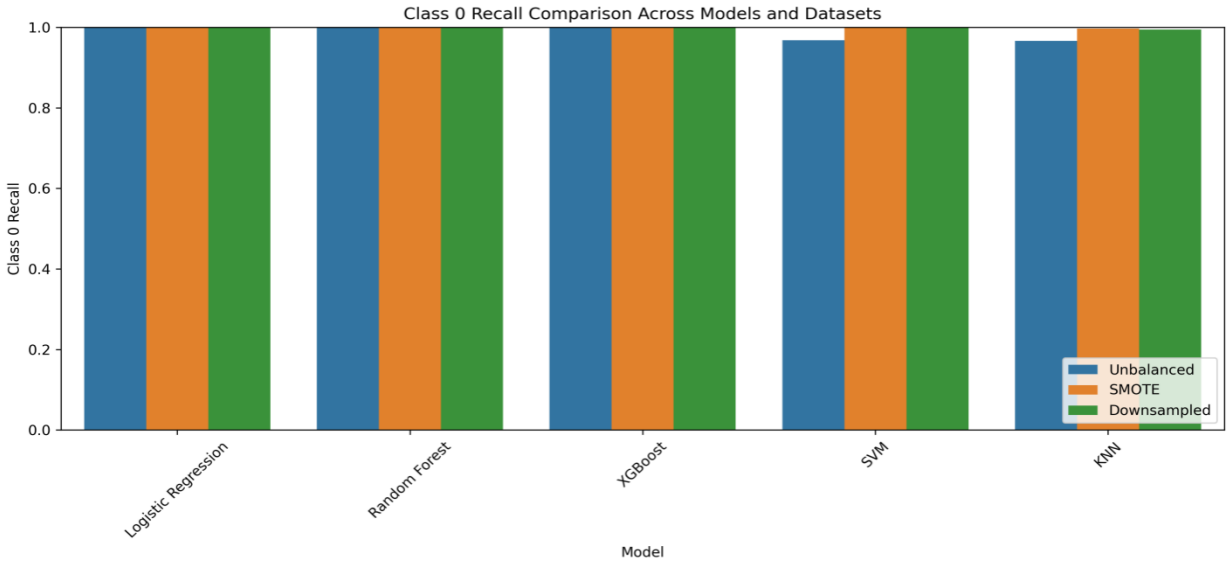

Figure 3: Recall Metric Performance of the Models on Fraudulent Transaction Data

Table 5 presents the performance results for the legitimate transactions (Class 1). The results show that the performance across all models was generally high due to the natural majority representation of this class. On the unbalanced dataset, Logistic Regression, Random Forest, and XGBoost maintained extremely high precision, recall, and F1-scores, exceeding 99.9%. Although SVM and KNN achieved slightly lower F1-scores of 99.12% and 99.01% respectively, their performance remained high. Under SMOTE, ensemble models and Logistic Regression sustained their superior results, while SVM demonstrated a somewhat unstable behaviour: although it achieved perfect precision, its recall declined to 95.87%, suggesting possible overfitting or conservative decision thresholds when exposed to synthetic data. KNN also performed well under SMOTE, attaining 99.72% precision and 97.89% recall. In the downsampled dataset, ensemble models again preserved near-optimal performance for the majority class, whereas SVM continued showing high precision but reduced recall, resulting in slightly lower F1-scores compared to other models. The precision and recall performance metric on the datasets across all the algorithms are presented in Figure 4 and Figure 5 respectively.

Table 5: Performance Evaluation Result on Class 1 (Legitimate Transaction) Prediction

Dataset	Model	Class 1 Precision	Class 1 Recall	Class 1 F1	Class 1 Support
Imbalanced	Logistic Regression	1.0000	0.9999	0.9999	8246
	Random Forest	0.9999	0.9999	0.9999	8246
	XGBoost	0.9995	0.9994	0.9995	8246
	SVM	0.9889	0.9935	0.9912	8246
	KNN	0.9883	0.9920	0.9901	8246
SMOTE	Logistic Regression	0.9998	0.9999	0.9998	8252
	Random Forest	0.9999	0.9999	0.9999	8252
	XGBoost	0.9990	0.9994	0.9992	8252
	SVM	1.0000	0.9587	0.9789	8252
	KNN	0.9972	0.9789	0.9880	8252
Downsampled	Logistic Regression	1.0000	0.9997	0.9998	3008
	Random Forest	1.0000	0.9987	0.9993	3008
	XGBoost	0.9990	0.9983	0.9987	3008
	SVM	1.0000	0.9398	0.9690	3008
	KNN	0.9942	0.9678	0.9808	3008

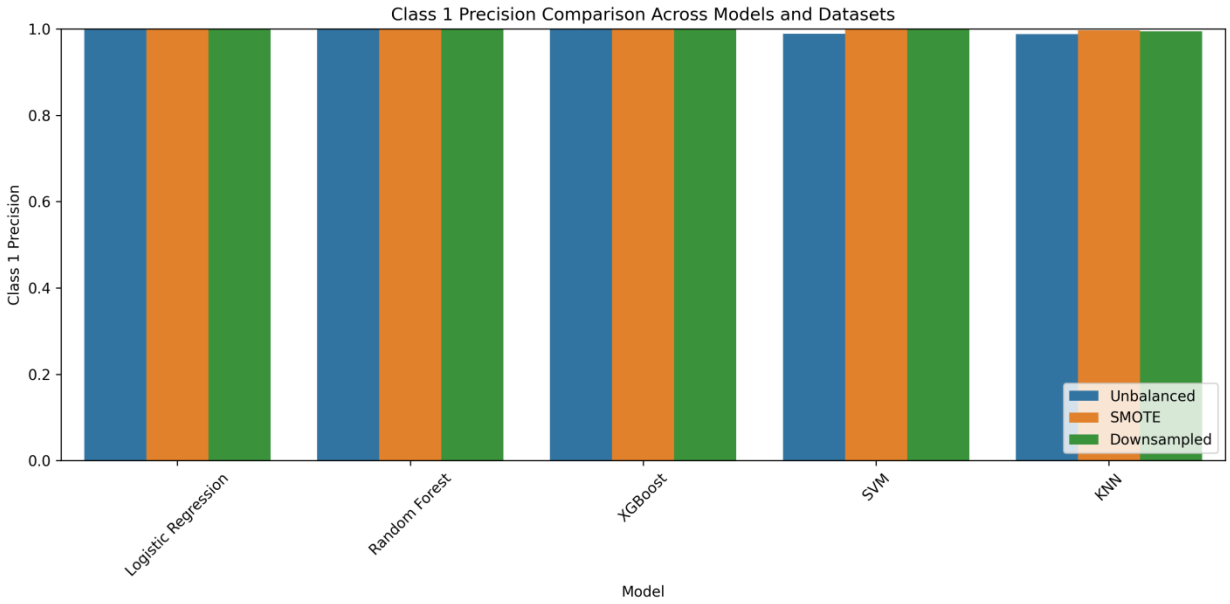

Figure 4: Precision Metric Performance of the Models on Legitimate Transaction Data

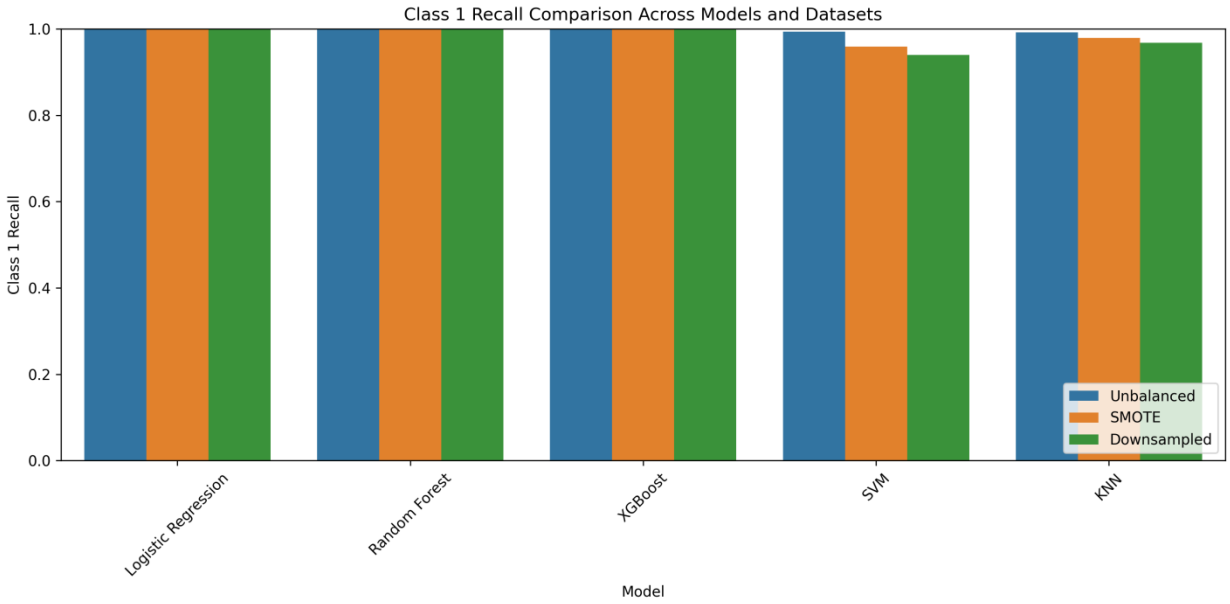

Figure 5: Recall Metric Performance of the Models on Legitimate Transaction Data

The overall model performances presented in Table 6, summarized through accuracy, average precision, average recall, F1-score, and AUC-ROC, further emphasize the superior stability of ensemble models. Across all datasets, Random Forest and XGBoost consistently achieved almost perfect accuracy, F1-scores, and AUC-ROC values of 1.000, illustrating excellent discrimination ability between fraud and legitimate transactions. Logistic Regression, despite being a simpler linear model, also delivered nearly equivalent performance, positioning it as a strong interpretable baseline. In contrast, SVM and KNN showed clear sensitivity to class imbalance. On unbalanced data, their F1-scores dropped to approximately 98%, but SMOTE substantially improved their results, raising both recall and precision closer to the levels observed for the ensemble methods. Downsampling led to similar improvements but remained less effective than SMOTE for these two algorithms.

Table 6: Overall Accuracy, AUC-ROC and Average Precision and Recall and F1 for two classes

Dataset	Model	Accuracy	Average Precision	Average Recall	Average F1	AUC-ROC
Imbalanced	Logistic Regression	0.9999	0.9998	0.9999	0.9999	1.0000
	Random Forest	0.9998	0.9998	0.9998	0.9998	1.0000
	XGBoost	0.9992	0.9989	0.9990	0.9989	1.0000
	SVM	0.9869	0.9850	0.9808	0.9828	0.9991
	KNN	0.9854	0.9826	0.9792	0.9809	0.9983
SMOTE-balanced	Logistic Regression	0.9998	0.9998	0.9998	0.9998	1.0000
	Random Forest	0.9999	0.9999	0.9999	0.9999	1.0000
	XGBoost	0.9992	0.9992	0.9992	0.9992	1.0000
	SVM	0.9792	0.9800	0.9793	0.9792	0.9999
	KNN	0.9880	0.9881	0.9880	0.9880	0.9977
Down sampled	Logistic Regression	0.9998	0.9998	0.9998	0.9998	1.0000
	Random Forest	0.9993	0.9993	0.9993	0.9993	1.0000
	XGBoost	0.9986	0.9986	0.9986	0.9986	1.0000
	SVM	0.9690	0.9699	0.9699	0.9690	0.9994
	KNN	0.9805	0.9804	0.9809	0.9805	0.9972

Collectively, these findings suggest that while class imbalance remains a significant challenge for fraud detection tasks, ensemble methods like Random Forest and XGBoost exhibit inherent resilience to imbalance and deliver superior performance without extensive data manipulation. SMOTE proves valuable for enhancing performance of algorithms like SVM and KNN that are otherwise more vulnerable to imbalance effects. The simplicity and stability of Logistic Regression further highlight its usefulness as a reliable baseline, particularly when model interpretability is desirable alongside high predictive accuracy. Overall, Random Forest and XGBoost are particularly well-suited for high-stakes domains such as financial fraud detection.

Conclusion

This study presented a comprehensive evaluation of multiple machine learning models for fraud detection using real-world bank card transaction data from Nigeria. The dataset exhibited significant class imbalance, with legitimate transactions vastly outnumbering fraudulent ones. Three data balancing strategies — unbalanced, SMOTE oversampling, and random downsampling — were applied to assess their influence on model performance. Five widely used classifiers (Logistic Regression, Random Forest, XGBoost, Support Vector Machine, and K-Nearest Neighbors) were trained and evaluated under each scenario.

The results demonstrate that ensemble models, particularly Random Forest and XGBoost, consistently achieved near-perfect performance across all metrics and datasets, displaying strong robustness to class imbalance. Logistic Regression also performed remarkably well, offering a highly interpretable yet accurate baseline. In contrast, SVM and KNN exhibited notable sensitivity to class imbalance but showed significant improvements when balanced

datasets were employed, especially through SMOTE oversampling. These findings reinforce the importance of selecting appropriate data balancing techniques, particularly for models that are less robust to skewed class distributions. Overall, ensemble models emerged as the most reliable choice for fraud detection tasks in highly imbalanced financial data settings.

Future Work

While this study has demonstrated the strong potential of traditional machine learning models for fraud detection, several promising directions remain for future investigation. First, more advanced class imbalance handling methods, such as cost-sensitive learning, adaptive synthetic sampling (ADASYN), or hybrid approaches, may further enhance performance, especially for models sensitive to imbalance. Second, integrating additional transactional and contextual information — such as temporal behavior, sequence patterns, and customer profiles — may capture complex fraud behaviors beyond isolated transactions.

An important future direction is the application of Graph Neural Networks (GNNs) to fraud detection. Since financial transactions naturally form complex relational structures involving customers, cards, merchants, and transactions, representing these entities as graphs allows models to learn richer dependencies and uncover hidden fraud patterns through message passing across the network. GNN-based models have shown significant promise in domains with relational data and could potentially capture sophisticated fraud schemes such as coordinated or collusive fraud rings, which are often difficult for classical models to detect.

Acknowledgement

This work was supported by funding provided by the Tertiary Education Trust Fund (Tetfund) under the Nigeria Ministry of Education through the Institutional Based Research Intervention scheme (2024 IBRI).

References

Adewale, A. (2020). The impact of BVN on fraud prevention in Nigerian banks. *Journal Financial Security,* 15(3), 45-60.

Bashir, S. A., Doolan, D. C., & Petrovski, A. (2016). The effect of window length on accuracy of smartphone-based activity recognition. *IAENG International Journal of Computer Science,* 43(1).

Bashir, S., Doolan, D., & Petrovski, A. (2015). *Clusternn: A hybrid classification approach to mobile activity recognition.* Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia, 263–267.

Berkmans, T., J., & Karthick, S. (2022). *Credit card fraud detection with data sampling,* 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICPECTS56089.2022.10046729.

CBN. (2018). BVN and the elimination of ghost accounts. Central Bank of Nigeria Report.

Carcillo, F., Le Borgne, Y. A., Caelen, O., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences.*

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). Scarff: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion,* 41, 182–194.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization. *International Journal of Data Science and Analytics,* 5(4), 285–300.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over sampling Technique. *Journal of Artificial Intelligence Research,* 16, 321-357.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. xx–xx): IEEE.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications,* 41(10), 4915–4928.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems,* 29(8), 3784–3797.

Elmangoush, A. M. A., Hassan, H., Fadhl, A. A. M., & Alshrif, M. A. (2024). Credit card fraud detection using synthetic minority oversampling technique and deep learning technique. 455–458. https://doi.org/10.1109/atsip62566.2024.10638849

Enwe O.; Fraud in Nigeria's Fin Tech Services: A regulatory weak link? *SRJ Legal*. 2021. https://srjlegal.com/fraud-in-nigerias-fin-tech-services-a-regulatory-weak-link/

Farabi, S. F., Ro, M. P., Alam, Md. M., Hossan, Md. S., Ariful, Md., Islam, Md. R., Uddin, A., Bhuiyan, M., & Biswas, M. N. (2024). Enhancing credit card fraud detection: A comprehensive study of machine learning algorithms and performance evaluation. *Journal of Business and Management Studies,* 6(3), 252–259. https://doi.org/10.32996/jbms.2024.6.13.21

Feng, X., & Kim, S.-K. (2024). Novel machine learning based credit card fraud detection systems. *Mathematics,* 12(12), 1869. https://doi.org/10.3390/math12121869

Jain, Y., Rathore, CA. D. S., Johrawanshi, A., Maheshwari, A., Pandey, A., & Saxena, N. (2024). Machine learning approaches for identifying fraudulent banking transactions: A financial management perspective. 1903–1909. https://doi.org/10.1109/ictacs62700.2024.10841041

Jayabalan, M. (2024). Use of machine learning in credit card fraud detection. https://doi.org/10.2174/9789815079661124010010

Jyoti, Er., Jindal, C., Singh, N. D., & Saini, A. (2024). improving credit card fraud detection through stacking ensemble models and SMOTE-ENN for imbalanced datasets. 1642–1649. https://doi.org/10.1109/ictacs62700.2024.10840967

Khedkar, D. S., & Gupta, B. (2024). Credit card fraud detection using machine learning. *International Journal of Advanced Research in Science, Communication and Technology,* 47–59. https://doi.org/10.48175/ijarsct-22307

Lebichot, B., Le Borgne, Y.-A., He, L., Oblé, F., & Bontempi, G. (2019). Deep-learning domain adaptation techniques for credit cards fraud detection. *In Recent Advances in Big Data and Deep Learning (INNSBDDL 2019)* Springer, 78-88.

Manda, V. T., Kondapalli, D., Malla, A., Jyothi, N. M., & Charan, Y. S. (2024). Imbalanced data challenges and their resolution to improve fraud detection in credit card transactions. https://doi.org/10.21203/rs.3.rs-3962043/v1

Mosa, D. T., Sorour, S. E., Abohany, A. A., & Maghraby, F. A. (2024). CCFD: Efficient credit card fraud detection using meta-heuristic techniques and machine learning algorithms. *Mathematics,* 12(14), 2250. https://doi.org/10.3390/math12142250

Misol J. K., & Agbadua E. G. (2022) Analysing and mitigating the problem of internet and credit card fraud in Nigeria. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT),* 8(3), 225-235. Available at doi: https://doi.org/10.32628/CSEIT228374

Njoku, D. O., Iwuchukwu V. C., Jibiri, J. E., Ikwuazom C. T., Ofoegbu C. I., Nwokoma F. O. (2024) Machine learning approach for fraud detection system in financial institution: A web base application. *International Journal of Engineering Research And Development*, 20(4), 01-12.

Otorokpo, A., Okpor, M. D., Yoro, E. R., Brizimor, S., Ifiokor, A. M., Obasuyi, D., Odiakaose, C.C., Ojugo, A. A., Atuduhor, R., Akiakeme, E., Ako, R.E., & Geteloma, V. O. (2024): DaBO-BoostE: Enhanced data balancing via oversampling technique for a boosting ensemble in card fraud detection. *Journal of Advances in Mathematical & Computational Science.* 12(1), 45-66.

Penumala, S. (2024). A comparative study of sampling techniques for imbalanced credit card fraud detection. *International Journal For Science Technology And Engineering,* 12(7), 789–800. https://doi.org/10.22214/ijraset.2024.63637

Priatna, W., Purnomo, H. D., Sembiring, I., & Wellem, T. (2024). Integrating class imbalance solutions into fraud detection systems: A systematic literature review. 1–6. https://doi.org/10.1109/ictiia61827.2024.10761334

Shah, A. & Makwana Y. (2023). Credit card fraud detection. *Research Gate*. http://www.researchgate.net/publication/369857378_Credit_Card_Fraud_Detection

Umar, A., Bashir, S. A., Abdullahi, M. B., & Adebayo, O. S. (2019). Comparative study of various machine learning algorithms for tweet classification. *i-Manager's Journal on Computer Science,* 6(4),12.