Enhancing Financial Fraud Detection using XGBoost, LSTM, and KNN with SMOTE for Imbalanced Datasets

Godfred Antwi Koduah[1*]; Jinguo Lian[2]

Department of Mathematics and Statistics, University of Massachusetts Amherst
*Corresponding Author

Received: 08 May 2025/ Revised: 15 May 2025/ Accepted: 24 May 2025/ Published: 05-06-2025

***Abstract*—** *The surge in digital financial activity has led to increasingly sophisticated forms of fraud, creating serious challenges for financial institutions. One of the core obstacles in fraud detection is the substantial class imbalance present in transactional datasets, where fraudulent records represent a small minority. This study presents a robust machine learning framework that integrates the Synthetic Minority Over-sampling Technique (SMOTE) with three distinct classifiers—XGBoost, Long Short-Term Memory (LSTM), and K-Nearest Neighbors (KNN)—to enhance the detection of fraudulent activities. Using a real-world dataset of six million banking transactions, we assess each model's performance through accuracy, precision, recall, F1-score, and both PR and ROC AUC metrics. Our findings show that SMOTE significantly boosts model recall and AUC scores. Among the models, XGBoost consistently delivers superior results with near-perfect metrics, while KNN maximizes recall, albeit at a slight cost to precision. LSTM produces more moderate but stable performance. Visual diagnostics, such as ROC/PR curves and confusion matrices, further confirm the reliability of XGBoost when combined with SMOTE. Overall, the integration of data balancing with advanced classifiers proves to be a powerful approach for real- time fraud detection.*

***Keywords*— Financial fraud detection, imbalanced datasets, machine learning, XGBoost, SMOTE.**

I. INTRODUCTION

With the rapid evolution of digital banking, the financial industry faces a mounting threat from fraudulent transactions. Fraud not only leads to significant monetary losses but also undermines consumer confidence in online financial platforms [1]. According to a 2022 report by the Association of Certified Fraud Examiners (ACFE), global fraud resulted in losses exceeding $3.6 billion, underscoring the urgent need for effective detection systems [2].

A major challenge in identifying fraudulent behavior lies in the highly skewed nature of fraud datasets, where valid transactions vastly outnumber fraudulent ones. This skewness often leads to machine learning models performing poorly on minority classes, resulting in high false negative rates and overlooked fraud [3].

To improve detection, a wide range of techniques has been investigated—ranging from rule-based heuristics to deep learning systems. While machine learning excels at identifying complex, non-linear patterns in high-dimensional data, its effectiveness is hindered by class imbalance [4]. To mitigate this, methods like the Synthetic Minority Over-sampling Technique (SMOTE) have been employed to create a more balanced training distribution by synthesizing new examples from the minority class [5].

In this study, we propose a data-centric approach to financial fraud detection that integrates SMOTE with three different learning models: XGBoost, LSTM, and KNN. These models were selected based on their complementary strengths: XGBoost for structured data classification, LSTM for temporal pattern recognition in transaction sequences, and KNN for localized anomaly detection. Our contributions are summarized as follows:

- **Synthetic Oversampling:** We apply SMOTE to increase the representation of fraudulent cases in the training data, improving recall and reducing bias toward the majority class.
- **Model Diversity:** We implement and compare three distinct algorithms—XGBoost, LSTM, and KNN—each optimized for specific aspects of the data.
- **Comprehensive Analysis:** We evaluate all models using six key metrics and visualize results through confusion matrices, ROC/PR curves, and histograms to provide a thorough performance assessment.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the dataset used in our experiments, including its features and preprocessing steps. Section 3 presents some machine learning models and data balancing methods. Section 4 provides evaluation and performance comparison of some machine learning models. Finally, Section 5 concludes the paper and suggests the best model for detecting fraud based on the results in section 4.

II. DATA DESCRIPTIONS

This study utilizes the Bank Account Fraud dataset, which is composed of six distinct subsets, each containing one million records, leading to a cumulative dataset of six million transaction entries. These include a base dataset along with five variant datasets, labeled I through V.

The base dataset, along with variants I, II, and IV, contains 32 features per transaction. Variants III and V expand slightly, including 34 features. The attributes across these datasets span a wide range of transaction characteristics, including:

- Demographic indicators (e.g., customer age, income level)
- Behavioral features (e.g., transaction frequency, session duration)
- Risk-related metrics (e.g., credit risk rating, proposed credit limits)
- Transactional data points (e.g., number of bank branches used, days since account activity)

A defining trait of the dataset is its significant class imbalance. The overwhelming majority of transactions are legitimate (class 0), while only a small fraction are labeled as fraudulent (class 1). Figure 1 in the paper presents a histogram illustrating this disproportion, emphasizing the difficulty of training fraud detection models on such imbalanced data.

A major challenge in fraud detection research is the limited availability of large-scale, real-world datasets, particularly those focused on new bank account (NBA) fraud. The dataset used here is one of the few publicly accessible and comprehensive datasets in this space. As such, it serves as a valuable benchmark for developing and evaluating the performance of machine learning models intended for fraud detection in the financial domain.

TABLE 1
DATA DESCRIPTION OF THE BASE DATASET

count	mean	std	min	25%	50%	75%	max	
fraud bool	1000000	0.01	0.1	0	0	0	0	1
income	1000000	0.56	0.29	0.1	0.3	0.6	0.8	0.9
name_email_similarity	1000000	0.49	0.29	0	0.23	0.49	0.76	1
prev_address_months_count	1000000	16.72	44.05	-1	-1	-1	12	383
current_address_months_count	1000000	86.59	88.41	-1	19	52	130	428
customer_age	1000000	33.69	12.03	10	20	30	40	90
days_since_request	1000000	1.03	5.38	0	0.01	0.02	0.03	78.46
intended balcon amount	1000000	8.66	20.24	-15.53	-1.18	-0.83	4.98	112.96
zip_count_4w	1000000	1572.69	1005.37	1	894	1263	1944	6700
velocity_6h	1000000	5665.3	3009.38	-170.6	3436.37	5319.77	7680.72	16715.57
velocity_ 24h	1000000	4769.78	1479.21	1300.31	3593.18	4749.92	5752.57	9506.9
velocity 4w	1000000	4856.32	919.84	2825.75	4268.37	4913.44	5488.08	6994.76
bank_branch_count_8w	1000000	184.36	459.63	0	1	9	25	2385
date_of_birth_distinct_emails_4w	1000000	9.5	5.03	0	6	9	13	39
credit risk score	1000000	130.99	69.68	-170	83	122	178	389
email is free	1000000	0.53	0.5	0	0	1	1	1
phone_home_valid	1000000	0.42	0.49	0	0	0	1	1
phone mobile valid	1000000	0.89	0.31	0	1	1	1	1
bank months count	1000000	10.84	12.12	-1	-1	5	25	32
has other cards	1000000	0.22	0.42	0	0	0	0	1
proposed_credit_limit	1000000	515.85	487.56	190	200	200	500	2100
foreign request	1000000	0.03	0.16	0	0	0	0	1
session_length_in_minutes	1000000	7.54	8.03	-1	3.1	5.11	8.87	85.9
keep alive session	1000000	0.58	0.49	0	0	1	1	1
device_distinct_emails_8w	1000000	1.02	0.18	-1	1	1	1	2
device_fraud_count	1000000	0	0	0	0	0	0	0
month	1000000	3.29	2.21	0	1	3	5	7

TABLE 2
DATA DESCRIPTION OF THE VARIANT I DATASET

count	mean	std	min	25%	50%	75%	max	
fraud bool	1000000	0.01	0.1	0	0	0	0	1
income	1000000	0,56	0.29	0.1	0.3	0.6	0.8	0.9
name_email_similarity	1000000	0.49	0.29	0	0.23	0.49	0.76	1
prev_address_months_count	1000000	16.96	43.87	-1	-1	-1	15	399
current address months count	1000000	83.59	86.46	-1	18	50	124	429
customer age	1000000	31.97	10.9	10	20	30	40	90
days since request	1000000	1.05	5.46	0	0.01	0.02	0.03	76.64
intended balcon amount	1000000	8.72	20.21	-15.74	-1.18	-0.83	6.22	113.12
zip count 4w	1000000	1574.5	1003.7	1	893	1270	1952	6678
velocity_6h	1000000	5661.9	3010.9	-174.11	3431.2	5300	7692.3	16817.8
velocity_ 24h	1000000	4767.1	1481.6	1322.3	3587	4745.6	5753.2	9539.36
velocity_4w	1000000	4857.2	919.76	2855.2	4269.2	4913.8	5488.6	7019.2
bank _branch_count_8w	1000000	181.17	457.64	0	1	9	24	2386
date _of_birth_distinct_emails_4w	1000000	9.86	5	0	6	9	13	39
credit risk score	1000000	129.41	69.07	-191	82	121	176	388
email is free	1000000	0.53	0.5	0	0	1	1	1
phone home valid	1000000	0.4	0.49	0	0	0	1	1
phone mobile valid	1000000	0.9	0.3	0	1	1	1	1
bank months count	1000000	10.8	12.12	-1	-1	5	25	32
has other cards	1000000	0.22	0.41	0	0	0	0	1
proposed_credit_limit	1000000	507.16	481.46	190	200	200	500	2100
foreign request	1000000	0.03	0.16	0	0	0	0	1
session length in minutes	1000000	7.46	7.95	-1	3.09	5.08	8.76	85.57
keep alive session	1000000	0.58	0.49	0	0	1	1	1
device distinctemails 8w	1000000	1.02	0.18	-1	1	1	1	2
device fraud count	1000000	0	0	0	0	0	0	0
month	1000000	3.29	2.21	0	1	3	5	7

TABLE 3
DATA DESCRIPTION OF VARIANT II DATASET

Column 01	count	mean	std	min	25%	50%	75%	max
fraud_bool	1000000.0	0.01	0.1	0.0	0.0	0.0	0.0	1.0
income	1000000.0	0.57	0.29	0.1	0.3	0.6	0.8	0.9
name_email_similarity	1000000.0	0.49	0.29	0.0	0.21	0.49	0.75	1.0
prev_address_months_count	1000000.0	14.82	43.23	-1.0	-1.0	-1.0	-1.0	399.0
current_address_months_count	1000000.0	99.38	94.56	-1.0	26.0	64.0	154.0	429.0
customer_age	1000000.0	41.3	13.8	10.0	30.0	50.0	50.0	90.0
days since_request	1000000.0	0.91	4.99	0.0	0.01	0.02	0.03	76.58
intended_balcon_amount	1000000.0	8.64	20.57	-15.74	-1.18	-0.83	0.08	112.7
zip_count_4w	1000000.0	1567.4	1009.62	1.0	901.0	1236.0	1909.0	6650.0
velocity_6h	1000000.0	5685.1	3001.71	-174.11	3470.24	5408.43	7653.99	16801.34
velocity_24h	1000000.0	4787.41	1470.37	1322.33	3628.56	4765.97	5750.78	9539.36
velocity_4w	1000000.0	4860.39	916.81	2870.59	4271.19	4919.35	5489.47	7019.2
bank_branch_count_8w	1000000.0	202.46	474.13	0.0	1.0	10.0	32.0	2377.0
date of birth distinct emails_4w	1000000.0	7.95	4.96	0.0	4.0	7.0	11.0	39.0
credit risk_score	1000000.0	137.46	72.2	-191.0	87.0	128.0	187.0	388.0
email_is_free	1000000.0	0.52	0.5	0.0	0.0	1.0	1.0	1.0
phone_home_valid	1000000.0	0.49	0.5	0.0	0.0	0.0	1.0	1.0
phone_mobile_valid	1000000.0	0.86	0.35	0.0	1.0	1.0	1.0	1.0
bank_months_count	1000000.0	11.2	12.11	-1.0	1.0	6.0	25.0	32.0
has_other_cards	1000000.0	0.24	0.43	0.0	0.0	0.0	0.0	1.0
proposed credit_limit	1000000.0	558.75	513.85	190.0	200.0	200.0	1000.0	2100.0
foreign_request	1000000.0	0.02	0.16	0.0	0.0	0.0	0.0	1.0
session_length_in_minutes	1000000.0	7.91	8.34	-1.0	3.21	5.28	9.42	87.24
keep alive_session	1000000.0	0.56	0.5	0.0	0.0	1.0	1.0	1.0
device_distinct_emails_8w	1000000.0	1.02	0.2	-1.0	1.0	1.0	1.0	2.0
device_fraud count	1000000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
month	1000000.0	3.29	2.21	0.0	1.0	3.0	5.0	7.0

TABLE 4
DATA DESCRIPTION OF VARIANT III DATASET

Column 01	count	mean	std	min	25%	50%	75%	max
fraud_bool	1000000.0	0.01	0.1	0.0	0.0	0.0	0.0	1.0
income	1000000.0	0.58	0.29	0.1	0.3	0.6	0.8	0.9
name_email_similarity	1000000.0	0.49	0.29	0.0	0.21	0.49	0.75	1.0
prev_address_months_count	1000000.0	14.67	43.02	-1.0	-1.0	-1.0	-1.0	399.0
current_address_months_count	1000000.0	99.23	94.07	-1.0	27.0	64.0	154.0	429.0
customer_age	1000000.0	41.34	13.77	10.0	30.0	50.0	50.0	90.0
days_since_request	1000000.0	0.9	5.01	0.0	0.01	0.02	0.03	76.58
intended balcon_amount	1000000.0	8.55	20.52	-15.74	-1.18	-0.83	-0.07	112.7
zip_count_4w	1000000.0	1517.66	965.03	1.0	886.0	1208.0	1844.0	6650.0
velocity_6h	1000000.0	5489.73	2940.94	-174.11	3332.99	5188.16	7367.06	16754.96
velocity 24h	1000000.0	4660.88	1451.48	1322.33	3503.01	4640.4	5591.86	9539.36
velocity 4w	1000000.0	4733.57	871.23	2870.59	4238.23	4813.0	5331.57	7019.2
bank_branch_count_8w	1000000.0	201.15	473.59	0.0	1.0	10.0	31.0	2377.0
date of birth distinct emails 4w	1000000.0	7.77	4.82	0.0	4.0	7.0	11.0	39.0
credit risk score	1000000.0	139.29	71.43	-177.0	90.0	130.0	188.0	388.0
email_is_free	1000000.0	0.52	0.5	0.0	0.0	1.0	1.0	1.0
phone_home_valid	1000000.0	0.49	0.5	0.0	0.0	0.0	1.0	1.0
phone mobile valid	1000000.0	0.86	0.35	0.0	1.0	1.0	1.0	1.0
bank months_count	1000000.0	11.14	12.13	-1.0	1.0	6.0	25.0	32.0
has_other_cards	1000000.0	0.25	0.43	0.0	0.0	0.0	0.0	1.0
proposed_credit_limit	1000000.0	551.69	506.66	190.0	200.0	200.0	1000.0	2100.0
foreign_request	1000000.0	0.02	0.15	0.0	0.0	0.0	0.0	1.0
session_length_in_minutes	1000000.0	7.81	8.23	-1.0	3.15	5.25	9.37	85.57
keep alive session	1000000.0	0.56	0.5	0.0	0.0	1.0	1.0	1.0
device_distinct emails_8w	1000000.0	1.02	0.19	-1.0	1.0	1.0	1.0	2.0
device_fraud_count	1000000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
month	1000000.0	3.66	2.12	0.0	2.0	4.0	5.0	7.0
x1	1000000.0	0.01	1.01	-4.98	-0.67	0.01	0.69	6.43
x2	1000000.0	0.01	1.01	-4.85	-0.67	0.01	0.68	6.54

TABLE 5
DATA DESCRIPTION OF VARIANT IV DATASET

	count	mean	std	min	25%	50%	75%	max
fraud_bool	1000000.00	0.01	0.10	0.00	0.00	0.00	0.00	1.00
income	1000000.00	0.58	0.29	0.10	0.30	0.60	0.80	0.90
name_email_similarity	1000000.00	0.49	0.29	0.00	0.21	0.49	0.75	1.00
prev_address_months_count	1000000.00	14.68	43.01	-1.00	-1.00	-1.00	-1.00	399.00
current_address_months count	1000000.00	99.21	94.08	-1.00	27.00	64.00	154.00	429.00
customer_age	1000000.00	41.34	13.78	10.00	30.00	50.00	50.00	90.00
days_since_request	1000000.00	0.90	5.01	0.00	0.01	0.02	0.03	76.58
intended_balcon_amount	1000000.00	8.55	20.52	-15.74	-1.18	8	-0.07	112.70
zip_count_4w	1000000.00	1517.55	964.96	1.00	886.00	1208.00	1844.00	6650.00
velocity_6h	1000000.00	5489.69	2940.44	-174.11	3333.59	5188.38	7366.62	16754.96
velocity_24h	1000000.00	4660.86	1451.39	1322.33	3502.92	4640.63	5591.77	9539.36
velocity_4w	1000000.00	4733.55	871.21	2870.59	4238.22	4813.07	5331.50	7019.20
bank branch count 8w	1000000.00	201.00	473.48	0.00	1.00	10.00	31.00	2377.00
date_of_birth_distinct_emails_4w	1000000.00	7.78	4.82	0.00	4.00	7.00	11.00	39.00
credit_risk_score	1000000.00	139.30	71.45	177.00	90.00	130.00	188.00	388.00
email is free	1000000.00	0.52	0.50	0.00	0.00	1.00	1.00	1.00
phone_home_valid	1000000.00	0.49	0.50	0.00	0.00	0.00	1.00	1.00
phone_mobile_valid	1000000.00	0.86	0.35	0.00	1.00	1.00	1.00	1.00
bank_months_count	1000000.00	11.14	12.13	-1.00	1.00	6.00	25.00	32.00
has_other_cards	1000000.00	0.25	0.43	0.00	0.00	0.00	0.00	1.00
proposed_credit_limit	1000000.00	551.73	506.71	190.00	200.00	200.00	1000.00	2100.00
foreign_request	1000000.00	0.02	0.15	0.00	0.00	0.00	0.00	1.00
session_length_in_minutes	1000000.00	7.81	8.23	-1.00	3.15	5.25	9.37	87.24
keep_alive_session	1000000.00	0.56	0.50	0.00	0.00	1.00	1.00	1.00
device distinct emails 8w	1000000.00	1.02	0.19	-1.00	1.00	1.00	1.00	2.00
device_fraud_count	1000000.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
month	1000000.00	3.66	2.12	0.00	2.00	4.00	5.00	7.00

TABLE 6
DATA DESCRIPTION OF VARIANT V DATASET

	count	mean	std	min	25%	50%	75%	max
fraud_bool	1000000.00	0.01	0.10	0.00	0.00	0.00	0.00	1.00
income	1000000.00	0.58	0.29	0.10	0.30	0.60	0.80	0.90
name_email_similarity	1000000.00	0.49	0.29	0.00	0.21	0.49	0.75	1.00
prev_address_months_count	1000000.00	14.74	43.13	-1.00	-1.00	-1.00	-1.00	384.00
current address months count	1000000.00	99.19	94.07	-1.00	27.00	64.00	154.00	426.00
customer_age	1000000.00	41.35	13.75	10.00	30.00	50.00	50.00	90.00
days_since_request	1000000.00	0.92	5.07	0.00	0.01	0.02	0.03	77.85
intended_balcon_amount	1000000.00	8.57	20.54	-15.71	-1.18	-0.83	-5	113.09
zip_count_4w	1000000.00	1517.47	965.95	1.00	885.00	1208.00	1846.00	6830.00
velocity_6h	1000000.00	5490.94	2940.12	-143.65	3332.98	5190.72	7371.56	16802.05
velocity_24h	1000000.00	4661.53	1450.44	1297.72	3505.06	4641.57	5593.34	9585.10
velocity_4w	1000000.00	4733.51	870.56	2858.75	4238.38	4813.95	5331.50	7019.20
bank_branch_count_8w	1000000.00	201.08	473.74	0.00	1.00	10.00	31.00	2404.00
date_of_birth_distinct_emails_4w	1000000.00	7.77	4.82	0.00	4.00	7.00	11.00	37.00
credit_risk_score	1000000.00	139.30	71.43	-177.00	90.00	130.00	188.00	388.00
email_is_free	1000000.00	0.52	0.50	0.00	0.00	1.00	1.00	1.00
phone_home_valid	1000000.00	0.49	0.50	0.00	0.00	0.00	1.00	1.00
phone_mobile_valid	1000000.00	0.86	0.35	0.00	1.00	1.00	1.00	1.00
bank_months_count	1000000.00	11.14	12.12	-1.00	1.00	6.00	25.00	32.00
has_other_cards	1000000.00	0.25	0.43	0.00	0.00	0.00	0.00	1.00
proposed_credit_limit	1000000.00	551.04	506.53	190.00	200.00	200.00	1000.00	2100.00
foreign_request	1000000.00	0.02	0.15	0.00	0.00	0.00	0.00	1.00
session_length_in_minutes	1000000.00	7.82	8.26	-1.00	3.15	5.25	9.36	83.21
keep_alive_session	1000000.00	0.56	0.50	0.00	0.00	1.00	1.00	1.00
device_distinct_emails_8w	1000000.00	1.02	0.19	-1.00	1.00	1.00	1.00	2.00
device_fraud_count	1000000.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
month	1000000.00	3.66	2.12	0.00	2.00	4.00	5.00	7.00
X1	1000000.00	0.01	1.01	-4.98	-0.67	0.01	0.68	6.43
x2	1000000.00	0.01	1.01	-4.85	-0.67	0.00	0.68	6.54

The histogram below shows the fraud class distribution of the Bank Account Fraud dataset.

FIGURE 1: Fraud Class Distribution

III. MACHINE LEARNING MODELS AND DATA BALANCING METHODS

3.1 Machine Learning Models:

This section outlines the core models employed for fraud detection—XGBoost, LSTM, and KNN—along with the data balancing technique SMOTE used to mitigate class imbalance. Each model was selected based on its strengths in dealing with structured data, temporal dependencies, and anomaly patterns.

3.1.1 XGBoost:

XGBoost (Extreme Gradient Boosting) is a highly efficient implementation of the gradient boosting algorithm, known for its scalability and predictive accuracy. It constructs an ensemble of decision trees, where each successive tree is trained to correct

the errors of the previous ones. This additive training approach minimizes a loss function and optimizes model performance over time.

Key features of XGBoost include:

- Built-in regularization to prevent overfitting
- Native support for handling missing values
- High compatibility with structured/tabular data

Due to its robustness and precision, XGBoost is commonly adopted in fraud detection tasks, especially where speed and accuracy are critical. Important tuning parameters include the number of trees (estimators), learning rate, and tree depth.

3.1.2 LSTM:

Long Short-Term Memory (LSTM) networks are a specialized form of Recurrent Neural Networks (RNNs) designed to learn long-term dependencies in sequential data. LSTMs incorporate gated memory cells that control the flow of information, allowing the network to retain or forget past data as needed.

This architecture makes LSTM ideal for analyzing time-series data, such as transaction histories, where order and temporal patterns are essential. In fraud detection, LSTMs help capture subtle behavioral sequences that could indicate anomalous activity.

Critical hyperparameters include:

- Number of LSTMunits
- Number of hidden layers
- Dropout rate for regularization

3.1.3 KNN:

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning method used for classification and regression. It classifies a data point based on the majority label of its ’k’ closest neighbors in the feature space.

KNNis particularly useful in scenarios with imbalanced data, as it can detect localized anomalies—clusters of fraudulent transactions that deviate from the norm. It also requires no training phase, making it computationally simple, though costly at prediction time.

Key factors influencing KNN performance include:

- The value of k (set to 5 in this study)
- The choice of distance metric (e.g., Euclidean or Manhattan)
- Weighting schemes for neighbours (uniform or distance-based)

3.2 Data Balancing with SMOTE:

Due to the severe imbalance in the dataset, standard training processes would result in models heavily biased toward the dominant (non-fraud) class. To correct this, we apply the Synthetic Minority Over-sampling Technique (SMOTE).

SMOTE addresses class imbalance by generating new synthetic instances of the minority class. Rather than duplicating existing samples, it interpolates between neighboring minority instances to create plausible new examples. This method enhances the model’s exposure to fraudulent behavior during training, improving its ability to generalize and detect rare events.

3.3 Integration of SMOTE with Machine Learning Models:

Before model training, SMOTE is applied to the training set to ensure a balanced representation of both classes. This preprocessing step ensures that the models—XGBoost, LSTM, and KNN—learn from a more equitable sample distribution.

As shown in later sections, applying SMOTE results in significant improvements across multiple performance metrics, particularly recall, F1 score, and AUC values, all of which are critical for identifying fraud cases.

3.4 Summary:

This section introduced the core components of our fraud detection framework: three machine learning models—XGBoost, LSTM, and KNN—and the SMOTE oversampling method for addressing class imbalance. The combination of these tools serves as the foundation for the evaluation and analysis described in the next section.

IV. MODEL EVALUATION AND PERFORMANCE COMPARISON:

This section presents a comparative analysis of the three selected machine learning models—XGBoost, LSTM and KNN—evaluated both with and without SMOTE. The assessment is based on six key performance metrics. In addition, we include visual tools such as ROC and PR curves, histograms, and confusion matrices to provide deeper insight into each model's behavior.

4.1 Evaluation Metrics:

To quantify the models' performance, we employ the following evaluation criteria:

4.1.1 Accuracy:

Accuracy measures the ratio of total predictions that are correct. It is computed as:

$$\textbf{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}} \quad (1)$$

However, in highly skewed datasets, accuracy can be misleading as it may reflect the dominance of the majority class rather than true performance on the minority (fraud) class.

4.1.2 Precision:

Precision measures the proportion of predicted fraud cases that are actually fraudulent:

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

High precision is desirable when false positives are costly, such as in fraud investigations that demand manual follow-up.

4.1.3 F1 Score:

The F1 score is the harmonic mean of precision and recall, offering a balanced view of a model's performance:

$$F1\ Score = \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

In fraud detection, where both missed fraud (FN) and false alarms (FP) carry consequences, F1 is an important metric.

4.1.4 Recall:

Recall, captures the proportion of actual fraud cases correctly identified:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

A high recall is particularly important in fraud detection, as missing fraudulent cases can lead to significant financial and reputational losses.

4.1.5 PR AUC:

The Precision-Recall Area under the Curve (PR AUC) evaluates the trade-off between precision and recall across different thresholds. It is particularly useful in skewed datasets where traditional ROC curves might be less informative.

$$PR\ AUC = \int Precision(r)\ dr \quad (5)$$

Where r represents recall. A perfect model achieves a PR AUC of 1, while random classifiers typically score close to the class prevalence.

4.1.6 ROC AUC:

The Receiver Operating Characteristic Area under the Curve (ROC AUC) shows the model's ability to differentiate between the positive (fraud) and negative (non-fraud) classes at various thresholds. A higher ROC AUC indicates better classification performance.

4.2 Performance Comparison with and without SMOTE:

Table 7 summarizes the models' performance with and without the application of SMOTE. As observed, using SMOTE leads to substantial gains in recall, F1 score, PR AUC, and ROC AUC for all models.

TABLE 7
PERFORMANCE METRICS FOR XGBOOST, LSTM, AND KNN WITH AND WITHOUT SMOTE

Model	Accuracy	Precision	F1 Score	Recall	PR AUC	ROC AUC
XGBoost (without SMOTE)	0.9916	0.8127	0.4237	0.2865	0.55	0.64
XGBoost (with SMOTE)	0.9928	0.9943	0.9891	0.984	1	1
LSTM (without SMOTE)	0.9917	0.8627	0.4099	0.2688	0.57	0.63
LSTM (with SMOTE)	0.8949	0.7676	0.8617	0.9822	0.95	0.98
KNN (without SMOTE)	0.9905	0.9371	0.2201	0.1247	0.54	0.56
KNN (with SMOTE)	0.9489	0.8671	0.9288	1	0.98	0.99

KNN experiences the most dramatic improvement in recall and F1 score, indicating its increased ability to detect fraud once the dataset is balanced. XGBoost, when paired with SMOTE, achieves perfect AUC values (both ROC and PR) and maintains strong precision and recall. LSTM also benefits from SMOTE, particularly in recall, though its overall performance remains more moderate.

4.3 Visualizing Model Performance with and without SMOTE:

To better understand how each model behaves, we provide several forms of visualization.

4.3.1 Histograms:

Figures 2 and 3 display the distribution of evaluation metrics before and after SMOTE is applied. These visualizations make it evident that SMOTE leads to a notable boost in recall, F1 score, and AUC values, while having minimal impact on accuracy and precision.

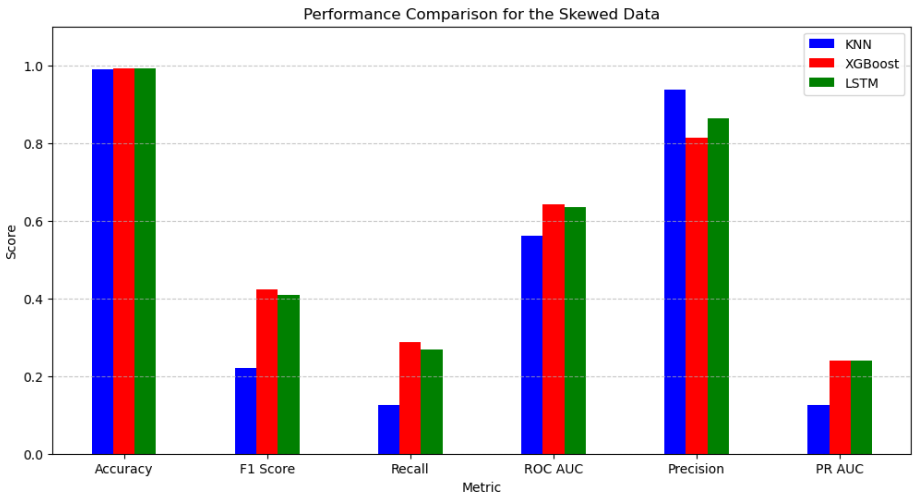

FIGURE 2: Distribution without SMOTE

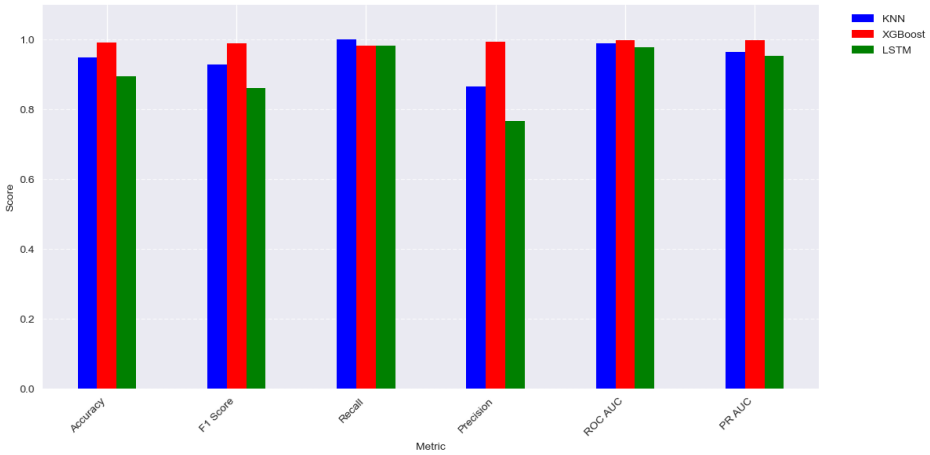

FIGURE 3: Distribution with SMOTE.

4.3.2 ROC Curves:

Figures 4 and 5 illustrate ROC curves for all three models under both conditions (with and without SMOTE). The curves confirm that balancing the dataset enhances the models' ability to differentiate between fraudulent and legitimate transactions, as reflected in higher AUC scores.

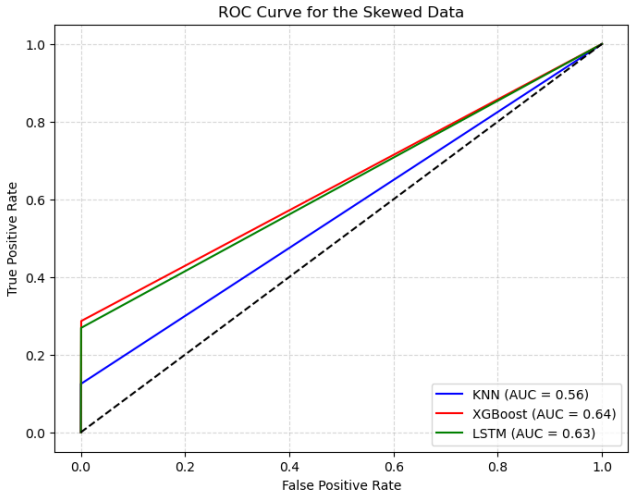

FIGURE 4: ROC curves without SMOTE

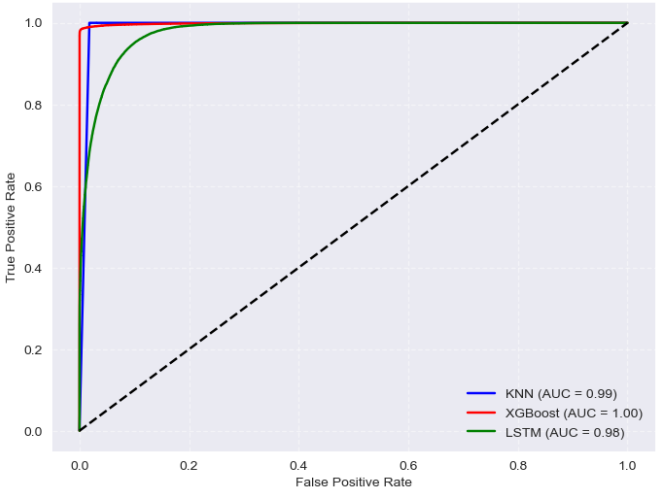

FIGURE 5: ROC curves with SMOTE

4.3.3 PR Curves:

Figures 6 and 7 present the Precision-Recall curves. These plots provide insight into the trade-offs models make as decision thresholds shift. Post-SMOTE results demonstrate improved recall while maintaining precision, particularly for XGBoost and LSTM.

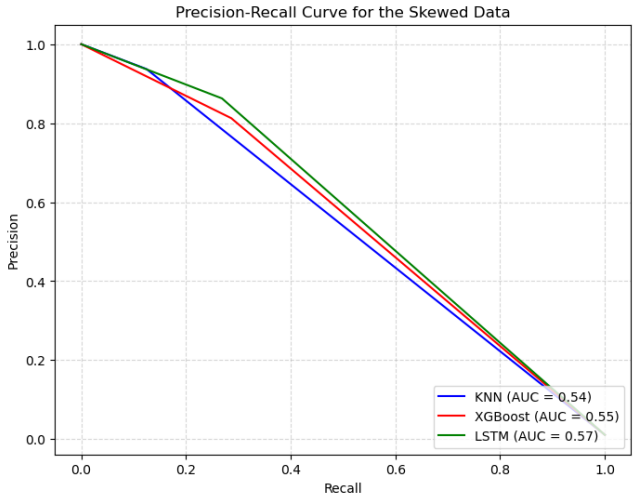

FIGURE 6: PR curves without SMOTE

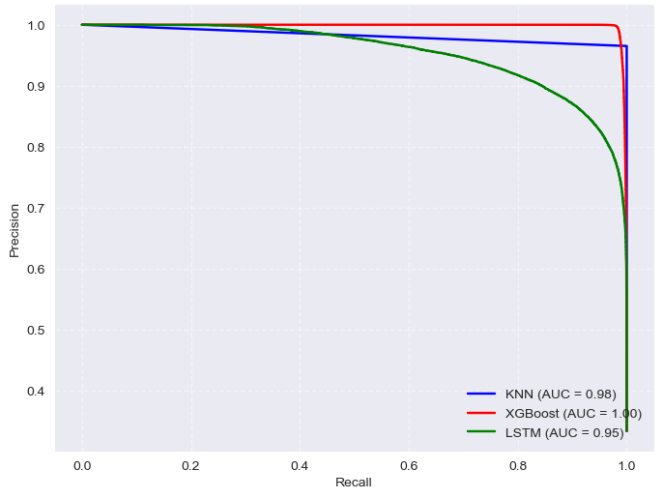

FIGURE 7: PR curves with SMOTE

4.3.4 Confusion Matrices:

Figures 8 and 9 show confusion matrices, offering a granular view of model classification results. After applying SMOTE, there is a visible reduction in false negatives (FN) across all models, especially in KNN, which shifts to identifying nearly all fraudulent transactions.

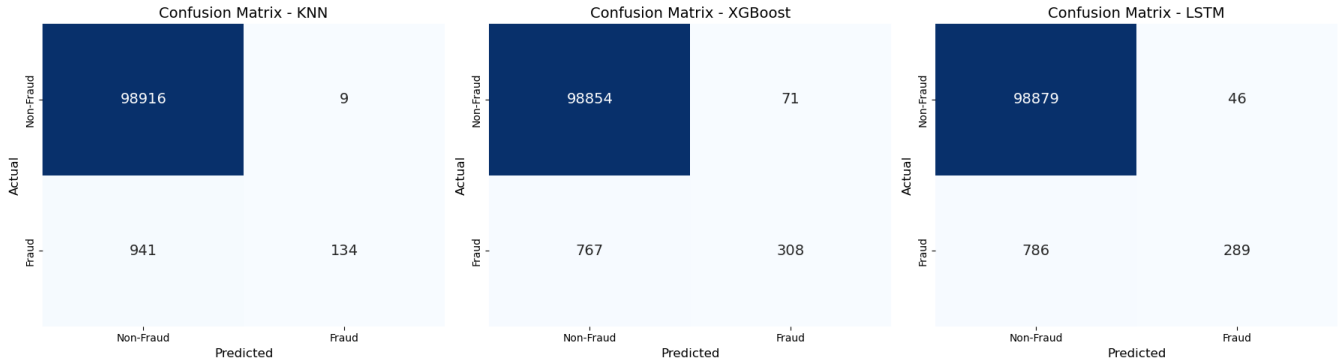

FIGURE 8: Confusion Matrix without SMOTE

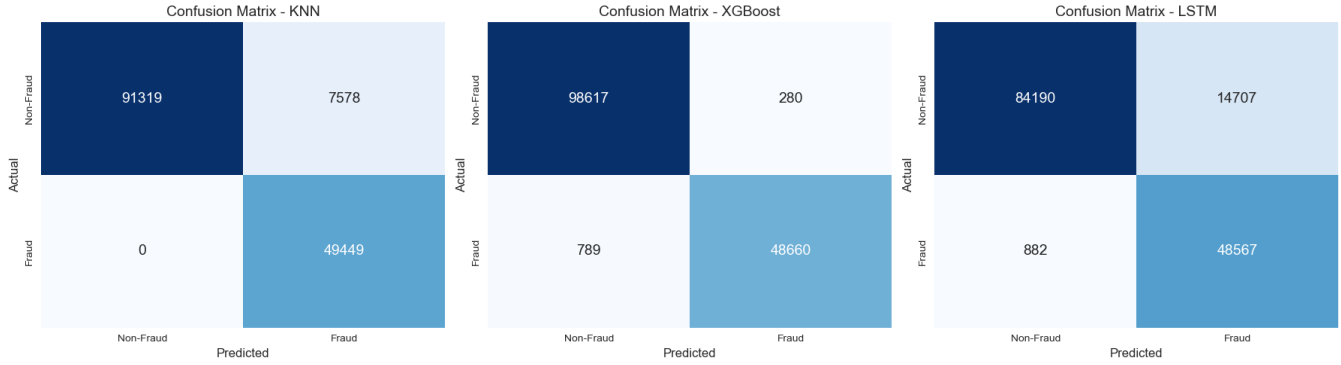

FIGURE 9: Confusion Matrix with SMOTE

4.4 Summary

The experiments confirm that SMOTE is an effective technique for improving fraud detection in imbalanced datasets. All three models benefit from its use, particularly in terms of recall and F1 score. XGBoost remains the top-performing model, delivering high precision and near-perfect AUC scores. KNN, while achieving perfect recall, sacrifices precision and suffers from higher false positives. LSTM shows a good balance but does not match XGBoost in overall performance.

V. CONCLUSION

This study explored the application of three machine learning models—XGBoost, KNN, and LSTM—for detecting fraudulent financial transactions, with a particular focus on addressing class imbalance through the use of SMOTE.

Our findings reveal clear differences in model behavior both before and after applying SMOTE. Without balancing, all models—especially KNN and LSTM—struggled with recall due to the scarcity of fraudulent instances in the training data. Among the unbalanced results, XGBoost stood out with the fewest false negatives, but still suffered from reduced sensitivity overall.

The introduction of SMOTE significantly improved each model's ability to detect fraudulent transactions. Recall increased across the board, most notably for KNN (from 0.1247 to 1.0000) and LSTM (from0.2688to0.9822), as evidenced by confusion matrices. However, these gains were accompanied by a rise in false positives, particularly for KNN, reflecting the classic precision-recall trade-off that arises in imbalanced classification problems.

Despite these trade-offs, XGBoost with SMOTE consistently emerged as the best-performing model, achieving outstanding results across all key metrics. It reached perfect PR AUC and ROC AUC scores (1.00) and maintained a strong balance between precision (0.9943) and recall (0.9840). While KNN achieved flawless recall, it did so at the cost of precision (0.8671) and overall F1 score stability. LSTM demonstrated considerable improvement post-SMOTE, but still lagged behind in precision and balanced performance.

These results were further reinforced by visual tools such as ROC and PR curves, histograms, and confusion matrices. The consistency of XGBoost's dominance across all metrics and visual diagnostics makes it a robust and scalable solution for real-world fraud detection systems, especially when augmented with data balancing techniques like SMOTE.

SOURCE OF DATA

The Bank Account Fraud data used in this research is publicly available at https://github.com/feedzai/bank-account-fraud.

REFERENCES

[1] T. Ashfaq et al., "A machine learning and blockchain based efficient fraud detection mechanism," *Sensors*, vol. 22, no. 19, p. 7162, Sep. 2022.

[2] ACFE, "Association of Certified Fraud Examiners (ACFE) 2022 Report to the Nations," 2022. [Online]. Avail- able: https://legacy.acfe.com/report-to-the-nations/2022/[Accessed: 2023]; E. Eber- lein et al., "Mathematics in Financial Risk Management," *Research Gate*, vol. 4, no. 1, pp. 1–26, 2007.

[3] N. S. Alfaiz and S. M. Fati, "Enhanced credit card fraud detection model using machine learning," *Electronics*, vol. 11, no. 4, p. 662, 2022.

[4] P. Vanini et al., "Online payment fraud: From anomaly detection to risk management," *Financial Innovation*, vol. 9, no. 1, p. 66, Mar. 2023, doi: 10.1186/s40854-023-00470-w.

[5] D. Gorton, "Modeling fraud prevention of online services using incident response trees and value at risk," in *Proc. 10th Int. Conf. Availability, Reliability and Security*, Toulouse, France, Aug. 2015, pp. 149–158, doi: 10.1109/ARES.2015.17.