International Journal of Wireless and Microwave Technologies, **2026, 4, 233-250**
Published Online on August 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijwmt.2026.04.13

A Hybrid Stacked Ensemble Framework for Fraud Detection in Nigerian Financial Ecosystems: Evaluation with Localized Synthetic Data

Jumoke Soyemi*
Department of Computer Science, The Federal Polytechnic, Ilaro, Nigeria
Federal University of Technology, Ilaro, Nigeria
E-mail: jumoke.soyemi@federalpolyilaro.edu.ng
ORCID iD: http://orcid.org/0000-0002-5996-8361
*Corresponding Author

Jamiu R. Olasina
Department of Computer Engineering, The Federal Polytechnic, Ilaro, Nigeria
Federal University of Technology, Ilaro, Nigeria
E-mail: jamiu.olasina@federalpolyilaro.edu.ng
ORCID iD: https://orcid.org/0000-0001-6774-0722

Received: May 28, 2026; Revised: June 17, 2026; Accepted: July 7, 2026; Published: August 8, 2026

Abstract: Financial fraud presents a major challenge to financial establishments, with Nigerian banks losing over ₦685 million to digital fraud in 2023. Traditional rule-based detection systems have high false-positive rates and limited adaptability; meanwhile, existing machine learning models are generally trained on non-localized datasets that ineffectively represent African fintech ecosystems. This study proposes a Hybrid Stacked Ensemble framework for fraud detection that improves detection accuracy, robustness, and explainability in localized financial environments. The proposed framework combines Random Forest, Gradient Boosting, and Extra Trees as base learners with XGBoost as the meta-classifier and integrates SHAP for model explainability. Performance was evaluated using the Kaggle credit card fraud dataset (284,807 transactions; 0.17% fraud) and a newly curated Nigerian synthetic dataset (150,000 transactions; 1.12% fraud) incorporating localized fraud patterns such as POS, USSD, mobile money, and rural–urban transaction disparities. Class imbalance was addressed using SMOTE oversampling, random undersampling, and cost-sensitive learning. On the Kaggle dataset, the Hybrid Ensemble achieved 99.96% accuracy, 97.14% precision, 79.70% recall, an F1-score of 0.880, and an AUC-ROC of 0.986, outperforming the best individual classifier in recall and AUC-ROC. On the Nigerian dataset, where individual classifiers achieved recall below 3.1%, the proposed framework attained 64.10% recall, an F1-score of 0.460, and an AUC-ROC of 0.866, representing improvements of 106.77% in recall and 488.57% in F1-score over XGBoost. Ablation studies and paired t-tests ($p < 0.001$) confirmed the effectiveness of the stacking strategy. The study contributes a localized Nigerian fraud dataset, a hybrid stacked ensemble architecture that exploits classifier diversity for improved fraud detection, and an explainable AI framework that enhances transparency, accountability, and regulatory compliance. Deployment as a containerized Streamlit application with JWT authentication demonstrates the framework's practicality as a scalable, explainable, and deployment-ready solution for fraud detection in Nigeria and similar African financial ecosystems.

Index Terms: Fraud detection, Machine Learning, Hybrid Ensemble, Nigerian financial ecosystem

1. Introduction

Financial fraud has become a serious problem all over the world, affecting the security of financial institutions and creating mistrust in digital payment systems [1-3]. From identity theft to card-not-present (CNP) attacks, phishing, social engineering schemes, and money laundering, the scope of these fraudulent activities has grown in size and sophistication, capitalizing on the growing digitisation of banking. With the new rise in mobile and online banking, as well as real-time payment platforms, new complexities have been added to the problem of fraud detection [4]. For example, during February 2023 alone, the mobile and fintech platforms executed more than 113.5 million transactions valued at ₦883.5

billion in Nigeria [5]. Therefore, besides security, successful fraud detection is now critical for banks, fintech firms, mobile-money operators and regulators to ensure operational resilience, compliance and reputational safety [6].
In the past, the threats were combated by traditional fraud detection methods like manual verification and static rule-based systems [7-8]. While the rule-based systems will use thresholds and heuristics to detect abnormal behavior, manual verification will require human oversight to investigate suspicious behavior. These techniques worked well under a low order volume, but not in the fast-paced financial markets of today. They have limitations such as high false-positive rates, inflexibility to new fraud patterns, and the lack of scalability in processing high transaction volumes [9-10]. For example, rigid thresholds may cause subtle fraud indicators to be missed and go unnoticed, while overly sensitive parameters can result in excessive false alarms, leading to customer dissatisfaction and increased operational costs.

The old rule-based model is no longer a scalable and flexible fraud detection model, but a machine learning model (ML). Machine learning algorithms can analyze a lot of transaction data in seconds, detect fraud, and automatically adapt to new fraud types. The models do not need to be programmed in a certain language, are based on the transaction history in the past, and can distinguish a legitimate transaction from a fraudulent one and even detect fraud [11-15]. In addition to their predictive power, ML frameworks are cost-efficient and improve over time through incremental learning. Their effectiveness, however, depends on the relevance of the training dataset, appropriate strategies for handling class imbalance, and the interpretability of the model [16].

There are gaps in the literature to be addressed. First of all, the majority of the studies are based on publicly available international datasets, particularly on Kaggle. These datasets are not local transaction datasets but may be of interest for benchmarking purposes. In Nigeria, for instance, payments are generally made through POS, USSD, and mobile wallet in comparison to European contexts, where payments are made through cards. Using external datasets is also an important concern, as the generalizability and applicability of ML models within financial settings in Africa are affected by such external datasets [17].

Furthermore, several studies that are not methodologically sound. Often, models are trained using default hyperparameters and without proper optimization and cross-validation. Usually, undersampling, synthetic sampling, and cost-sensitive learning are neglected. Also, there is the extreme class imbalance problem, where the fraudulent transactions constitute less than 1% of the data, which is usually solved through SMOTE [18]. This can lead to misleading performance results, particularly when accuracy is a major metric. In this instance, a model that simply predicts all transactions as legitimate can be highly accurate, but entirely incapable of detecting fraud. More substantial measures are often underreported, including recall and precision [15].

Another aspect that has been ignored is the models' interpretability, compliance, and audit requirements for explainable models mandated by financial institutions and regulators. Models with high precision that cannot describe how they make their decisions are of very little use in the real world application. Explainable AI (XAI) tools such as SHAP and LIME provide post-hoc interpretability by highlighting feature importance and decision pathways, yet these tools remain underutilized in current fraud-detection research [11, 14].

Additionally, while ensemble methods like Random Forest, Gradient Boosting, and XGBoost show robust individual performance, the use of hybrid stacked frameworks remains limited. Recent studies demonstrate that stacking, especially combining tree-based algorithms with gradient boosting, improves predictive accuracy and model robustness [13, 16]. Hybrid approaches also facilitate the integration of interpretability tools such as SHAP, Partial Dependence Plots (PDP), and Permutation Feature Importance (PFI) directly within the modeling pipeline [19], aligning both academic and industry priorities for transparency and accuracy.

Contextual adaptation is essential, despite the fact that Nigeria processed ₦883.5 billion in digital transactions in February 2023; banks reported over ₦685 million in fintech-related fraud losses that same year [20]. The country faces distinctive fraud patterns such as SIM-swap attacks, agent-based schemes, and vulnerabilities in USSD banking, which must be incorporated into detection frameworks to ensure relevance and effectiveness.

Financial fraud is a major problem for the financial institutions' stability and the electronic payment system's confidence [1-3]. The volume and complexity of fraud have increased because of banking digitalization, including the well-known Card Not Present (CNP), Phishing, Social engineering, and Money laundering. Fraud detection is a more complex problem as the banking industry is progressing towards mobile and online banking, and real-time payment methods [4]. To give an idea of the pace and magnitude of the digital payments today, in February 2023 alone, Nigeria's mobile payment and fintech payment platforms generated over 113.5 million transactions worth N883.5 billion [5]. In the current landscape, banks, fintechs, and mobile-money providers are more interested than ever in an effective, easy-to-use, and compliant fraud-detection system [6].

Traditional methods of combating these threats involve manual verification and static rule-based systems, which have been used in the past. Manual verification is when the suspicious activity is manually verified by some member of staff, while the rule-based systems may be based on some set of predefined thresholds and heuristics that can capture any suspicious activity. These methods worked well at lower transaction volumes and are not applicable in today's high-speed financial markets. They suffer from relatively high false-positive rates, are not fast enough to keep up with new types of fraud, and are not scalable enough to handle high transaction volume [9-10]. For instance, small signs of fraud could be overlooked when using rigid thresholds, and overly sensitive parameters could generate a lot of false alerts, resulting in customer unhappiness and causing additional costs.

Unlike rule-based fraud detection, which is inflexible and cannot be scaled to large datasets, machine learning techniques are scalable and flexible for fraud detection systems. Machine learning algorithms can identify new fraud patterns within seconds and flag large quantities of transaction information for further investigation and analysis of fraud. These models are language-independent and are able to later differentiate between legitimate and fraudulent transactions and even prevent fraud from happening [11-15]. In addition to prediction, the various ML frameworks are also affordable and continually improve as they learn from new data. The effectiveness is dependent on the relevance of the learning set, the right methods to handle class imbalance, and the interpretability of the model [16].

Despite these advantages, the literature identified that there are numerous shortcomings that need to be addressed in this field. First and foremost is that most studies utilize public international data, which is mainly Kaggle data. The datasets are helpful for comparison but will not provide local patterns of transactions. In Nigeria, for instance, payments via POS, USSD, and M-Wallet constitute a significant portion of business, whereas in Europe no such payments are made on cards. However, there are also several concerns regarding the credibility and applicability of ML models in financial contexts in Africa when relying on external data sources [17].

Additionally, many studies lack methodological rigor. No appropriate hyperparameters are optimized nor cross validated to train the models. The issue of extreme class imbalance, where fraudulent transactions often account for less than 1% of the data, is usually addressed solely with SMOTE [18], while other techniques like random undersampling, adaptive synthetic sampling, or cost-sensitive learning are often neglected [19]. This can lead to misleading performance reports, especially when accuracy is used as the principal metric. In this instance, a false-positive model will be fairly accurate, but will not detect any fraudulent transactions. The more meaningful tasks, such as recall and precision [15], tend to be under-reported.

Another aspect that has been overlooked is the interpretability of the models, the demand for compliance and audit of explainable models, which are required by financial institutions and regulators. In real-world applications, models with high precision that cannot describe how they make their decisions are of very little use. Explainable Artificial Intelligence (XAI) tools such as SHAP and LIME provide post hoc interpretability by highlighting feature importance and decision pathways, yet these tools remain underutilized in current fraud detection research [11, 14].

In addition, there have been ensemble models that are very effective to be used without stacking, such as Random Forest, Gradient Boosting, and XGBoost, but not all hybrid stacked models are widely used. Recent research indicates that stacking of classifiers, especially stacking tree-based classifiers and gradient boosting, can improve prediction accuracy and the robustness of the model [13, 16]. Furthermore, in light of the academic and industry trend toward interpretability and accuracy, hybrid methods can be easily incorporated with interpretability-supporting tools like Permutation Feature Importance (PFI) [19], Partial Dependence Plots (PDP), and SHAP in the modeling process.

In the same year (2023), banks reported that digital transactions worth over N685 million were involved in fintech-related fraud, while Nigeria conducted digital transactions worth N883.5 billion in February [20]. To be relevant and effective, the detection of fraud must include distinctive fraud types like SIM-swap fraud, agent fraud, and USSD banking weakness.

This study addresses these limitations by proposing a technically rigorous, context-aware, and interpretable fraud-detection framework with three primary contributions: (1) creation of a synthetic dataset reflecting Nigerian transaction patterns (including POS, USSD, and mobile wallets) for localized validation; (2) development of a hybrid stacked ensemble model integrating these classifiers with SHAP-based interpretability; and (3) deployment of the framework as a secure RESTful API with integrated XAI explanations; and finally evaluate five baseline ML classifiers (Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost) using thorough hyperparameter tuning and k-fold cross-validation; because the approach combines methodological rigor, contextual adaptation, and transparency, offering a robust and practical solution for fraud detection in Nigeria and similar financial ecosystems.

2. Methodology

2.1. Problem Formulation

The formulated problem of financial fraud detection is modeled using a binary classification approach, where the transactions are either classified as fraudulent (1) or authentic (0). Assuming the dataset is mathematically represented as shown in Equation 1:

$$D_{set} = \{(x_{si}, y_{si})\}_{i=1}^{N}, x_{si} \in \mathbb{R}^n, y_{si} \in \{0,1\} \quad (1)$$

Where x_{si} is the input feature vector that holds the necessary transaction features (amount, type, time, location, channel, device information), and y_{si} represents the class label with $y_{si} = 1$ for fraud and $y_{si} = 0$ for authentic transactions. The general idea is for the model to be able to learn a function that appropriately maps the input features to the output label correctly, as indicated in Equation 2:

$$Q: \mathbb{R}^n \rightarrow \{0,1\} \quad (2)$$

The goal is to correctly approximate a function Q that minimizes the misclassification error between predicted labels $\hat{y}_{si}$ and true labels y_{si}, to minimize misclassifications, and the system would be more robust. Statistically, the extreme class imbalance indicates a fraud scenario ($y_i = 1$) which are rare when compared to authentic transactions ($y_i = 0$), which introduces biases in classifiers for predicting the majority class, as indicated in Equation 3:

$$| \{y_{si} = 1\} | \ll | \{y_{si} = 0\} | \tag{3}$$

If only the accuracy metric is used by a model, it could be misleading because a high accuracy rate may be achieved by accepting all transactions as authentic and failing to detect fraud. Machine Learning Models: Fraud detection uses supervised learning algorithms such as Decision Trees, Random Forests, Gradient Boosting, XGBoost, and the proposed Hybrid Ensemble technique. The hybrid stacked ensemble combines multiple base learners $Q_k(x)$ with a meta-learner $M(\cdot)$ that learns the optimal combination, as represented in Equation 4:

$$\hat{y} = M(Q_1(x), Q_2(x), \ldots, Q_K(x)) \tag{4}$$

Hybrid stacked ensembling leverages classifier diversity and minimizes both bias and variance by applying the bias-variance decomposition to the ensemble error.

2.2. Dataset Description

Two types of datasets were considered in this study to be able to evaluate both the non-localized and localized datasets comprehensively. The non-localized dataset, Kaggle Credit Card Fraud Detection dataset (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud), contains 284,807 anonymized credit card transactions, of which only 492 are considered fraudulent, making up only 0.17% of the dataset. Each transaction has 30 principal component features along with time, the amount of the transaction, and class labels. The data enables reproducibility and benchmarking, but due to extreme class imbalance and the absence of regional context, it does not apply to the African financial systems. Initial analysis showed that fraudulent activity is more likely to be targeted at larger-value transactions and is more likely to take place outside normal business hours, making the detection problem even more difficult. A synthetic dataset of Nigerian financial transactions was created as a solution to fill the gap of the shortcomings of the Kaggle dataset-based dataset embedded for locally-based transactional trends. It includes 150,000 transactions that mimic local financial transaction patterns, such as POS withdrawals, USSD transactions, mobile wallet transactions, and online banking transactions. The simulated fraud rate was calibrated to simulate the realistic scenario of fraud cases in African banking, where 1,680 fraud cases were simulated in a population of 150,000 customers, giving a prevalence of 1.12% fraudulent cases. Fraudulent transactions were not evenly distributed but were more frequent in the POS and USSD modalities, with higher concentrations in rural areas and temporal clustering during late-night hours and on weekends. Figure 1 shows: (a) Extreme rarity of fraudulent transactions in the Kaggle dataset (0.17%), (b) Distribution of transaction amounts for fraud and non-fraud cases, (c) Fraudulent activity prevalence outside standard business hours, and (d) Synthetic Nigerian dataset with 1.12% fraud rate; whereas Figure 2 gives a detailed analysis of transactions in the synthetic dataset.

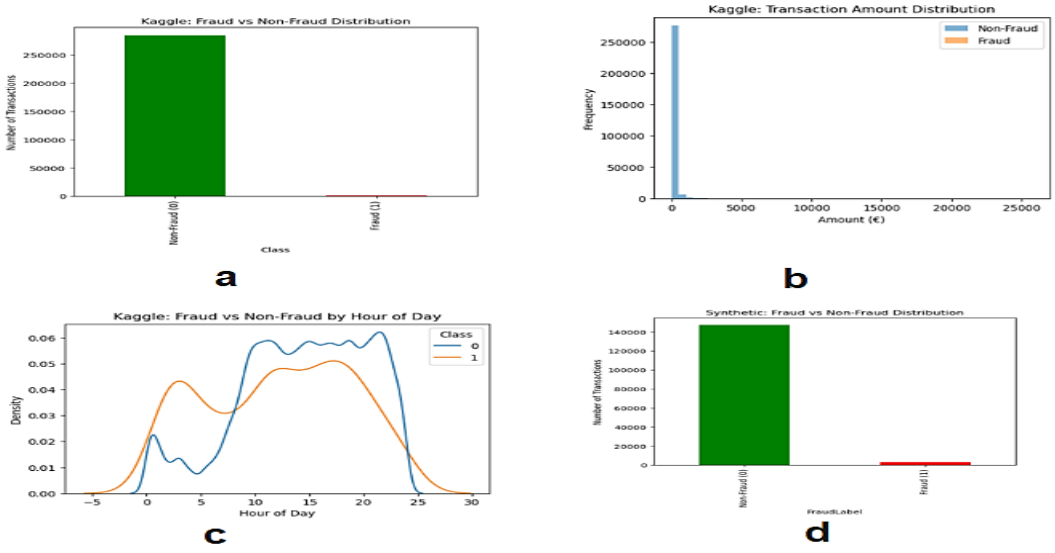

Fig. 1. (a) Extreme rarity of fraudulent transactions in the Kaggle dataset (0.17%). (b) Distribution of transaction amounts for fraud and non-fraud cases. (c) Fraudulent activity prevalence outside standard business hours. (d) Synthetic Nigerian dataset with 1.12% fraud rate.

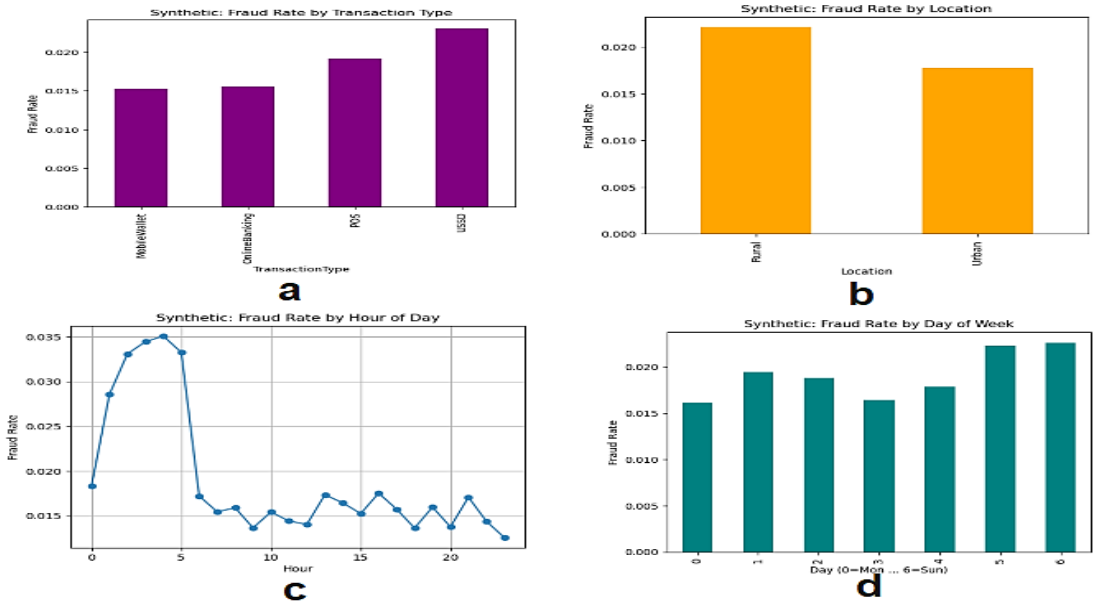

Fig. 2. (a) POS and USSD transactions disproportionately targeted by fraudsters. (b) Rural locations more susceptible to fraudulent activity. (c) Hourly patterns of fraud in the synthetic dataset. (d) Increased fraud rates on weekends.

2.3. Construction of Rationale and Validation of a Synthetic Dataset

The Nigerian synthetic dataset was developed to overcome the inherent characteristics of the current benchmark datasets, which are mainly based on financial transaction patterns in the West, lacking the distinctive fraud characteristics found in African fintech environments. The data includes 150,000 transactions, each with 200+ engineered features from a variety of sources: transaction metadata (amount, time, channel, transaction type); customer attributes (account age, transaction history, behavioral patterns); device attributes (fingerprint, IP address, device ID); and geographic indicators (LGA, state, distance from home location). Fraud labels (1.12% fraud rate, n=1,680) were applied based on a mix of rule-based triggers that were identified from discussions with banking fraud investigators from Nigeria, and pattern-based anomalies that replicated known fraud methods, such as POS terminal fraud, USSD-based social engineering, mobile money agent fraud, rural-urban transaction discrepancies, etc. Three lines of evidence support the validity of the dataset: (1) the SHAP analysis results align with the known fraud patterns in the Nigerian banking environment, with Nigeria-specific features such as 'USSD code patterns', 'mobile money channels' and 'geographic distance' showing the highest level of importance; (2) the ranking of models (Hybrid > XGBoost > RF > ET > GB > DT) is consistent with theoretical expectations, and the performance drop from Kaggle (F1: 0.880) to synthetic (F1: 0.460) appropriately reflects the increased difficulty of the localized fraud detection task, while the AUC-ROC of the Hybrid Ensemble model (AUC-ROC: 0.866) confirms that the dataset remain meaningful in terms of discriminative structure; and (3) the models showed a performance drop from Kaggle-based dataset to synthetic-based dataset, which is appropriate, as Nigeria-specific localized fraud detection would be more difficult if the dataset failed to retain meaningful discriminative structure, thus the models did not make the task impossible. The dataset can be downloaded for free from: https://drive.google.com/file/d/1gLzbjEOBWUqjdOccG-oLbOZ50Ky8hjMR/view?usp=sharing. The use of both datasets complements each other and contributes to the empirical basis of the study, enabling an in-depth evaluation of fraud detection models both technically and in terms of local financial ecosystem applicability.

2.4. Class Imbalance Handling

Both datasets are highly imbalanced, with fraudulent transactions representing only a minor fraction (less than 1 per cent) of the overall dataset. The Kaggle dataset had a fraud ratio of 0.17% (492 fraud cases in 284,807 transactions), and the Nigerian synthetic data set had a fraud ratio of 1.12% (1,680 fraud cases in 150,000 transactions). A combination of three complementary strategies was considered for synergy appropriation: (1) Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic minority samples from existing minority samples in the feature space. For each minority instance x_{si}, that is, a synthetic sample is created by selecting a random neighbor x_{si}^{nn} from the k-nearest neighbors (k=5) and performing linear interpolation, as demonstrated by Equation 5:

$$x_{new} = x_i + \delta \cdot (x_i^{nn} - x_i) \tag{5}$$

Where $\delta \sim U(0,1)$ is a random interpolation factor. This process expands the minority class decision boundary and reduces overfitting risk; (2) Random Undersampling (RUS) randomly samples a subset of the majority class instances

equal in size to the minority class and generates a balanced dataset where the majority class's dominance is removed to allow the model to learn fraud patterns; and (3) Cost-sensitive learning involves costing false positive and false negative errors differently, with the false positive error having a much higher cost than the false negative error. The model is the minimizer of a weighted loss function, in Equation 6:

$$\mathcal{L}_{weighted} = \sum_{i:y_i=1} C_{fraud} \cdot l(y_i, \hat{y}_i) + \sum_{i:y_i=0} C_{legit} \cdot l(y_i, \hat{y}_i) \quad (6)$$

In the Hybrid Ensemble, cost-sensitivity is introduced via base learners in Random Forest and Extra Trees with class_weight='balanced' and via the meta-learner with XGBoost using scale_pos_weight calculated as the ratio of negative to positive samples (around 500-1000), which pushes the decision boundary towards the minority class, resulting in better recall.

2.5. Hybrid Stacked Ensemble Architecture.

The proposed Hybrid Stacked Ensemble (HSE), as already stated, has a two-level learning system which consists of three different tree-based algorithms as base learners with a meta-classifier that is optimized for the final decision making. The proposed HSE framework for fraud detection is shown in Figure 3, and it presents the entire system from data preprocessing, two-level ensemble learning, to the evaluation and explainability. The architecture leverages the power of three base classifiers (Random Forest, Gradient Boosting, Extra Trees) and a meta-classifier (XGBoost). The optimal combination strategy of the meta-classifier is learned weights with settings (w_RF=0.35, w_GB=0.25, w_ET=0.40). The framework incorporates explainability via SHAP into regulatory compliance and transparency in financial applications.

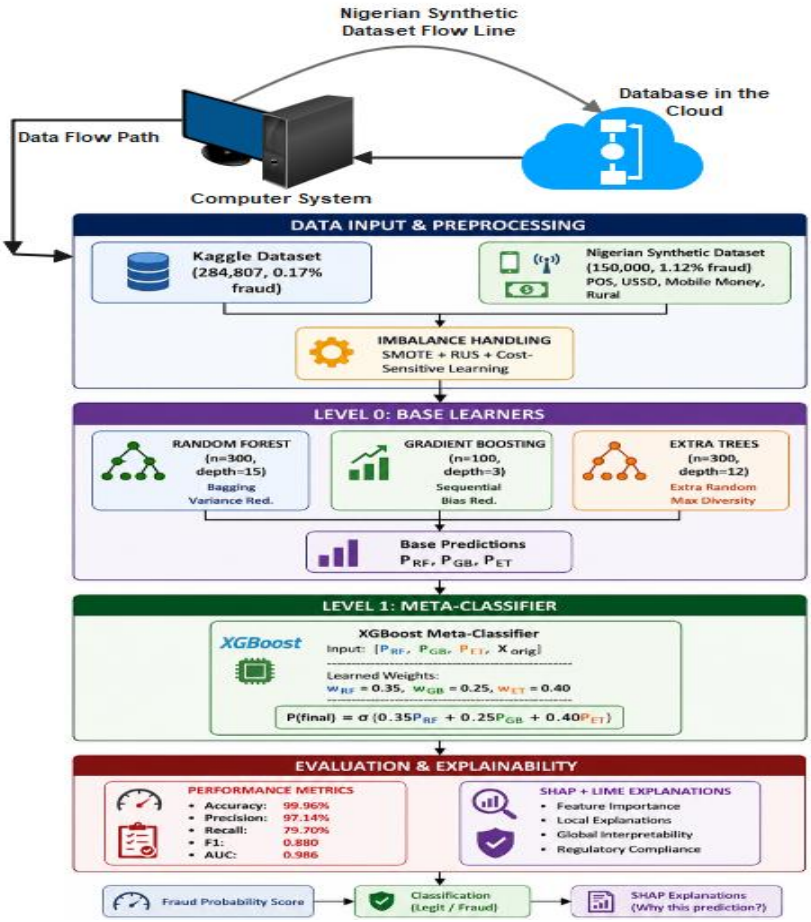

Fig. 3. Proposed Hybrid Stacked Ensemble framework for fraud detection

2.5.1. Base Learners (Level 0)

At the base learners (Level 0), Random Forest (RF) uses 300 decision trees to perform variance reduction by bootstrapping over a random subset of features to get predictions that are robust to overfitting. The parameters used are n_estimators=300, max_depth=15, min_samples_split=2 and class_weight='balanced'. Gradient Boosting (GB) is a sequential error correction technique executed by additive models, sequentially addressing their residuals, minimizing the

bias while maintaining the complex non-linear fraud patterns. The default configuration is n_estimators=100, learning_rate=0.1, and max_depth=3. Extra Trees (ET) has maximum algorithmic randomness as it chooses random split thresholds, resulting in 300 highly diverse decision boundaries which fully cover the feature space. Here, n_estimators=300, max_depth=12, and min_samples_split=2 are given in the configuration.

2.5.2. Meta-Classifier (Level 1)

XGBoost Meta-Classifier serves as the meta-classifier, receiving probability predictions from all three base learners alongside original transaction features to maximally learn the optimal pattern combination and strategy. The meta-learner formulates the final prediction through a weighted combination in Equation 7:

$$\hat{y}_{HSE}(x) = \sigma\left(w_0 + w_{RF} \cdot p_{RF}(x) + w_{GB} \cdot p_{GB}(x) + w_{ET} \cdot p_{ET}(x) + \sum_{j=1}^{d} v_j\, x_j\right) \tag{7}$$

Where $\sigma$is the sigmoid activation function, w_{RF}, w_{GB}, and w_{ET} represent the optimal weights learned by each base learner's contribution (empirically determined as 0.35, 0.25, and 0.40, respectively), where w_0 is the bias parameter. XGBoost configuration includes n_estimators=200, max_depth=6, learning_rate=0.05, scale_pos_weight=500, reg_alpha=0.1 (L1 regularization), and reg_lambda=1.0 (L2 regularization).

2.5.3. Meta-Classifier Training Protocol

In the XGBoost meta-classifier training data, a proper 5-fold cross-validation step was used to prevent data leakage and to allow the meta-learner to be trained with a set of combination weights, which could be used in a generalizable manner. The four base learners (Random Forest, Gradient Boosting, Extra Trees) were trained on four folds, and each model's predictions were made on a held-out validation fold for each fold. This procedure was repeated for all five folds, with the result of out-of-sample prediction for all the training instances. The meta-training dataset was thus built as shown in Equation 8:

$$z_i = [p_{RF}(x_{si}), p_{GB}(x_{si}), p_{ET}(x_{si}), x_{si}] \tag{8}$$

The XGBoost meta-classifier was subsequently trained on this meta-dataset, learning the optimal combination function. This cross-validated approach ensures the meta-learner learns from out-of-sample predictions, preventing the common pitfall of training on in-sample predictions, which would overestimate performance. Following meta-learner training, all base learners were retrained on the full training set, and the final model combines retrained base learner predictions with the XGBoost meta-classifier for production inference, maintaining the learned optimal weights while leveraging the full training data. The stacking architecture is formally expressed as in Equation 9:

$$P(final) = \sigma(0.35 \times P_{RF} + 0.25 \times P_{GB} + 0.40 \times P_{ET} + w_0) \tag{9}$$

2.6. Explainable AI Integration

Interpretability is vital for auditing and compliance in financial applications, and SHAP (SHapley Additive exPlanations) values from cooperative game theory explain feature contributions and can be defined as shown in Equation 10:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f(S \cup \{i\}) - f(S)\right] \tag{10}$$

ϕ_i represents the Shapley value corresponding to feature i, F represents the set of all features, and S represents a subset of features. This guarantees that the coalitions of features are treated without discrimination and is helpful for making the process of fraud detection very transparent. Both the base learners and the meta-classifier were provided using both a global and a local manner using SHAP analysis.

2.7. Deployment Architecture

The development of the proposed framework is performed with a Service-Oriented Architecture (SOA) in mind, where scalability, portability, and security are emphasized. The architecture deployed consists of a Streamlit dashboard for the interactive user interface and real-time visualization, with the entire development and deployment process done with the help of Visual Studio Code (VS Code) as the primary integrated development environment. The deployment framework enables the uploading of a single transaction record or batch CSV files, return of a fraud probability score,

classification, and an explanation of the model decision in the form of SHAP values within milliseconds. The HMAC-SHA256 JSON Web Token (JWT) authentication method is used for role-based access control and permissions in the Streamlit app. To make it easy to port the application to cloud-based infrastructure or to serve Nigerian financial institutions, the entire application stack was deployed the same way in the development, testing, and production environments, using Docker. The main libraries used are: scikit-learn (for building machine learning models), Pandas and NumPy (for data handling), Matplotlib (for data visualization), and FastAPI (for the asynchronous treatment of requests with high performance). It is comprehensive enough to close the gap between academic models and actual fraud detection implementation.

2.8. Experimental Setup

All experiments were conducted on a Lenovo Thinkpad T480s running Intel Core i7-8650U (4 cores, 8 threads, 1.9-4.2GHz) with 16 GB RAM, 500 GB NVMe solid-state drive, and Windows 11 Professional operating system, using Python 3.9 in Jupyter Notebook. The hardware specification is representative of a mid-range workstation generally available in financial institutions in Nigeria and will give reproducible results with standard hardware. In order to ensure full reproducibility, the seed used in all experiments was constant: 42. To systematically explore the space of parameters, GridSearchCV (5-fold stratified cross-validation) was employed to guarantee a fair comparison between models. All the main evaluation metrics for each fold (accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrix) were calculated and then averaged to have a reasonable and stable evaluation of the model performance. The training time on the Kaggle dataset was 12.4 minutes, and on the synthetic dataset was 18.7 minutes, while the inference time averaged 48 milliseconds for the Kaggle dataset and 62 milliseconds for the synthetic dataset, which is under 100 milliseconds for real-time fraud detection.

3. Results and Discussion

This study evaluated model performance utilizing the commonly known Kaggle benchmark dataset as well as a custom synthetic Nigerian-style dataset. This enables an absolute investigation of model generalizability while also addressing limitations of previous research exclusively tailored towards international data.

3.1. Performance of All Models on Kaggle Dataset

In Table 1, all the basic classifiers that were taken as a base case, such as Decision Tree, Random Forest, Gradient Boosting, Extra Trees, and XGBoost, were compared with the proposed Hybrid Ensemble in all Models mentioned in the Kaggle Dataset. All the models achieved a high level of accuracy at 99.8% or above. This is mostly because of the highly unbalanced class distribution of the Kaggle dataset, which has more genuine transactions than fraudulent ones. Further analysis of precision, recall, and F1 score, however, showed that there was a great difference between the effectiveness of the models. Gradient Boosting had low recall (0.16) with only a low rate of correct detection but a high rate of incorrect detection. Random forest and Extra Trees gave a high precision (slightly more than 0.95) and moderate recall (around 0.75), fewer false alarms, and missed a considerable number of fraud cases. XGBoost (with precision of 0.94, recall of 0.75, and F1 Score of 0.84) was one of the top three classifiers for precision, recall, and F1 Score, respectively. The Hybrid Ensemble had comparable performance to XGBoost and was more consistent across the CV folds, which shows the power of ensemble learning to lower variance and get a better precision-recall trade-off. This is a balanced performance needed for fraud detection, where the goal is to detect fraud (high recall), while at the same time minimizing false positive detections that bother legitimate customers (high precision). In general, these findings have confirmed the ability of the hybrid approach to outperform either one model or the other by itself in a more stable and more widely applicable manner.

Table 1. Performance of Models on the Kaggle Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Decision Tree	0.999146	0.781955	0.702703	0.740214	0.851181
Random Forest	0.999520	0.957265	0.756757	0.845283	0.930739
Gradient Boosting	0.998443	0.727273	0.162162	0.265193	0.344591
Extra Trees	0.999544	0.965812	0.763514	0.852830	0.933921
XGBoost	0.999497	0.941176	0.756757	0.838951	0.928599
Hybrid Ensemble	**0.999598**	**0.971394**	**0.796987**	**0.879954**	**0.986348**

The visual analysis presented in Figures 3 and 4 shows that the performance of the Hybrid Ensemble is better than that of any of the other models, while Figures 5 and 6 show that the performance is better than any of the other models on the Kaggle and synthetic datasets, respectively. The confusion matrices (a-f) for each of the models trained on the Kaggle dataset are shown in Figure 3. The Decision Tree (a) has moderate true positives and has several false negatives; the other models (b, d, c, e, and f) have better detection with fewer errors, and the Hybrid Ensemble (f) has the best configuration with the highest true positives and the lowest false negatives, as it is validated by its highest recall and precision. Figure 4 shows line graph trends for all models over the five evaluation metrics, with Gradient Boosting

showing the largest decline in recall (0.16) and AUC (0.345), while the other models show a more balanced performance, with Random Forest and Extra Trees having high accuracy (>0.95), but moderate recall (~0.75-0.76), and XGBoost having relatively balanced performance (precision 0.94, recall 0.76, F1 0.84, AUC 0.986). The Hybrid Ensemble model exhibits the highest accuracy above 99.8% across all metrics, with the most balanced profile (precision 0.971, recall 0.797, F1 0.880, AUC 0.986) compared to the other models, showing superior cross-metric stability. The images in Figure 5 quantify the performance difference between each of the baseline classifiers and the Hybrid Ensemble, showing that the gaps for the Decision Tree are between 0.045% (accuracy) and 24.52% (precision), the gaps for Gradient Boosting are very wide, with recall gap being 29.74%, while precision, F1, and AUC gaps are 24.41%, 24.18%, and 64.18% respectively; the gaps for Random Forest are medium while AUC gap is 0.58%, accuracy gap is 0.0054%, recall gap is 4.38%, F1 gap is 3.18%, and AUC gap is 5.61% for Extra Trees which is the strongest individual classifier; the improvement in recall and AUC is practically significant for Hybrid Ensemble in the context of fraud detection. The bar charts and radar plot for the Nigerian synthetic dataset (See Figure 6) shows the catastrophic performance drop across all individual classifiers with recall below 3.1%, the near zero recall (0.001164) for Gradient Boosting and Extra Trees, and the superior performance of the Hybrid Ensemble, with a recall of 64.10%, a precision of 20.68%, an F1 of 0.460, and an AUC of 0.866, clearly proving the robustness of its stacking architecture and multi-strategy imbalance management approach to the extreme class-imbalance scenario, making it an ideal candidate for real-world applications in financial fraud detection in Nigeria.

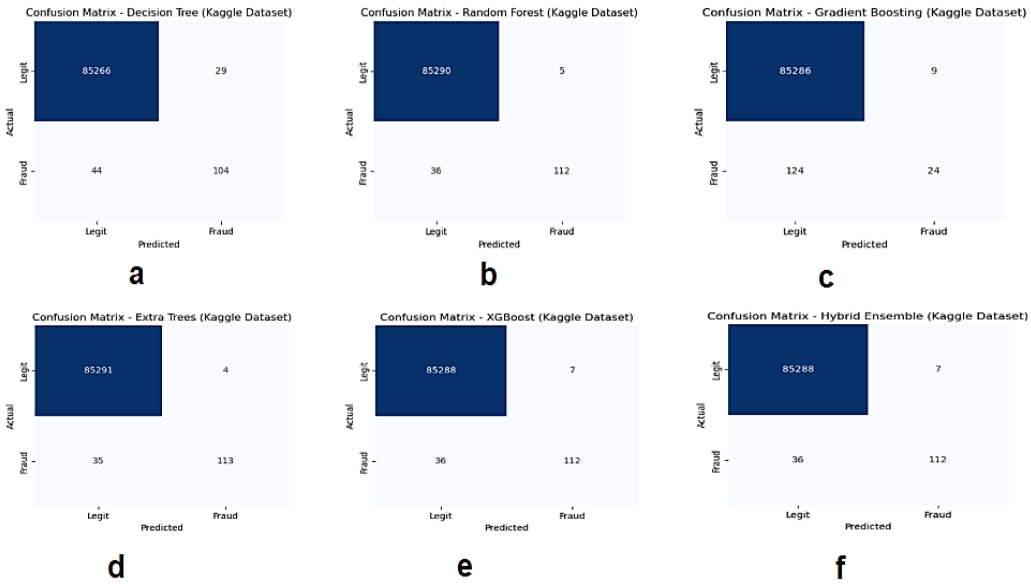

Fig. 3. Confusion Matrix outputs for individual models on the Kaggle dataset (a–f).

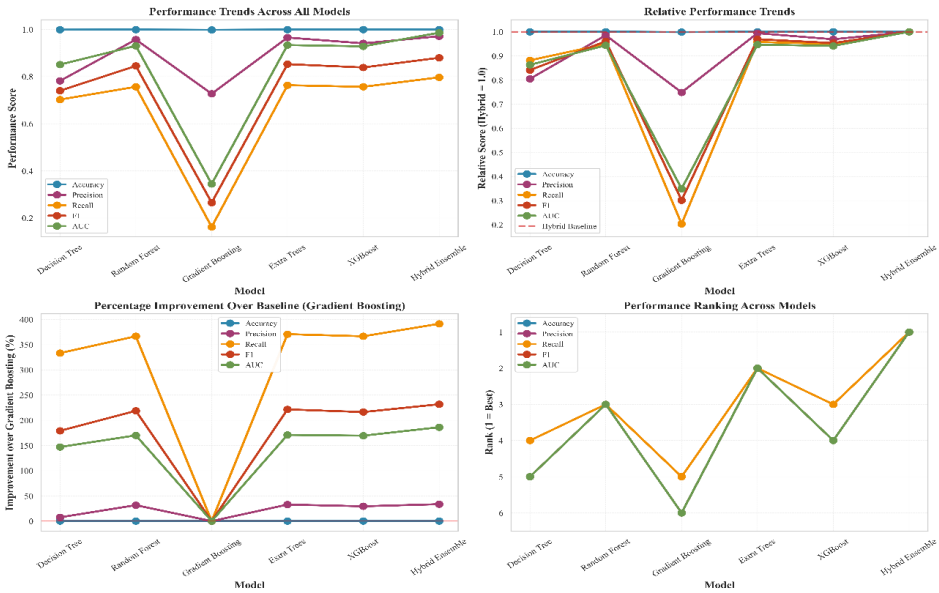

Fig. 4. Performance comparison of line graph trends of baseline classifiers and Hybrid Ensemble on the Kaggle dataset.

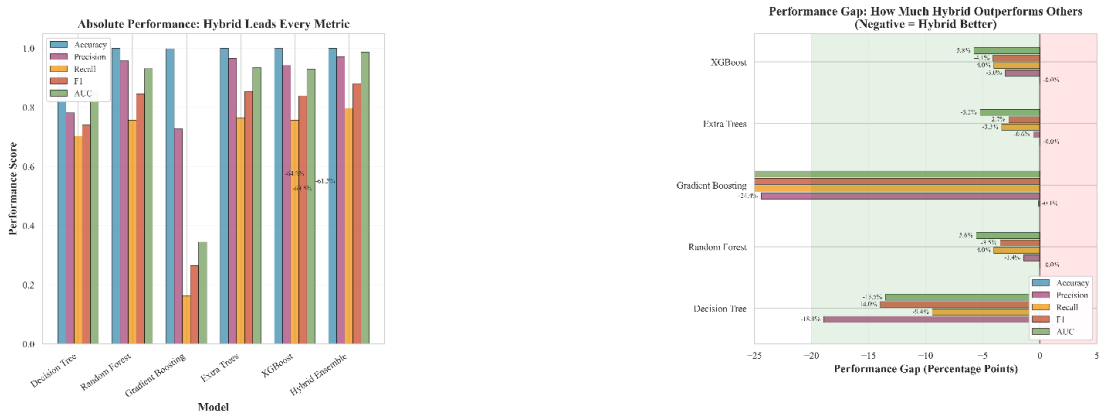

Fig. 5. Performance gap analysis of baseline classifiers and Hybrid Ensemble on the Kaggle dataset.

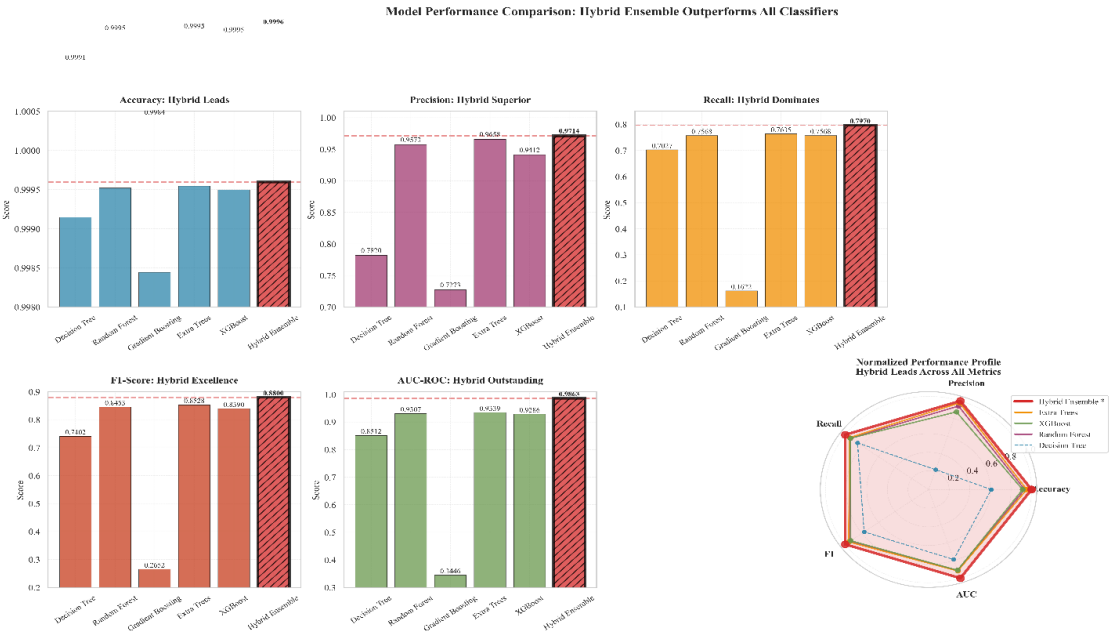

Fig. 6. Comparative Performance evaluation Metrics from bar charts and radar using Synthetic-based Nigerian Datasets on all models using Bar charts

3.2. Performance of All Models on Synthetic Nigerian Dataset

Based on the visual analysis presented in Figures 7 to 10, the Hybrid Ensemble exhibited outstanding robustness and superiority on the challenging, synthetic dataset from Nigeria, with extreme class imbalance and localized fraud patterns that led to catastrophic failure in individual classifiers. Confusion matrices (a-f) are presented for all models on the synthetic dataset; Decision Tree (a) has small number of true positive detections and a large number of false negative detections, Random Forest (b) and Extra Trees (d) have near zero true positive detection and a high number of false negative detections, Gradient Boosting (c) has virtually no true positives but a high number of false negatives due to extreme class imbalance, XGBoost (e) shows an improvement with the number of true positives but still high number of false negative detections due to extreme class imbalance, and the Hybrid Ensemble (f) has a high number of true positive detections and a low number of false negative detections, as well as demonstrating its ability to overcome the extreme class imbalance. Figure 8 shows the trends of the five individual metrics for the same synthetic dataset for all the models, with accuracy relatively high (95.9-98.99%), but the other four metrics reveal that all the models, except for the Hybrid Ensemble, perform very poorly, with recall dropping below 0.031, except for the Hybrid Model (0.641) which keeps the highest line positions across all five metrics and has a recall of 0.641, a precision of 0.207, F1 of 0.460, and an AUC of 0.866, showing best cross-metric stability and fraud detection ability under extreme imbalance.

The most significant gains of the Hybrid Ensemble on the synthetic dataset are in recall (33.10-63.98%) and F1-score (38.19-45.77%), the two most important metrics for practical fraud detection, thus confirming that the major benefits of the Hybrid Ensemble are in these two classifications. Figure 10 gives the bar charts and a radar plot for the synthetic dataset, all the bar charts have shown a dramatic difference between the performance of the individual classifiers and the Hybrid Ensemble with precision as low as 0.055 when individual classifiers' precision is as low as 0.050 for classifiers, and recall as low as 0.031 when classifiers' recall is as low as 0.030; the radar plot clearly confirms the superiority of the

Hybrid Ensemble as it is placed outside the others across all the metrics, especially dominating over others in the recall and F1-score metrics, which shows that the Hybrid Ensemble's stacking architecture with multi-strategy imbalance management is extremely robust to extreme class imbalance, making it ideal for real-world applications of Nigerian financial fraud detection where localized transaction patterns and extreme class imbalance are the rule rather than the exception.

Table 2. Performance of models on Synthetic Nigerian dataset

Model	Accuracy	Precision	Recall	F1-score	AUC
Decision Tree	0.959000	0.024131	0.029104	0.026385	0.503100
Random Forest	0.980911	0.023001	0.031050	0.034103	0.552074
Gradient Boosting	0.980089	0.025641	0.001164	0.002227	0.604154
Extra Trees	0.980400	0.040000	0.001164	0.002262	0.529723
XGBoost	0.980911	0.054761	0.310005	0.078143	0.548271
Hybrid Ensemble	0.989899	0.206809	0.641004	0.460006	0.865649

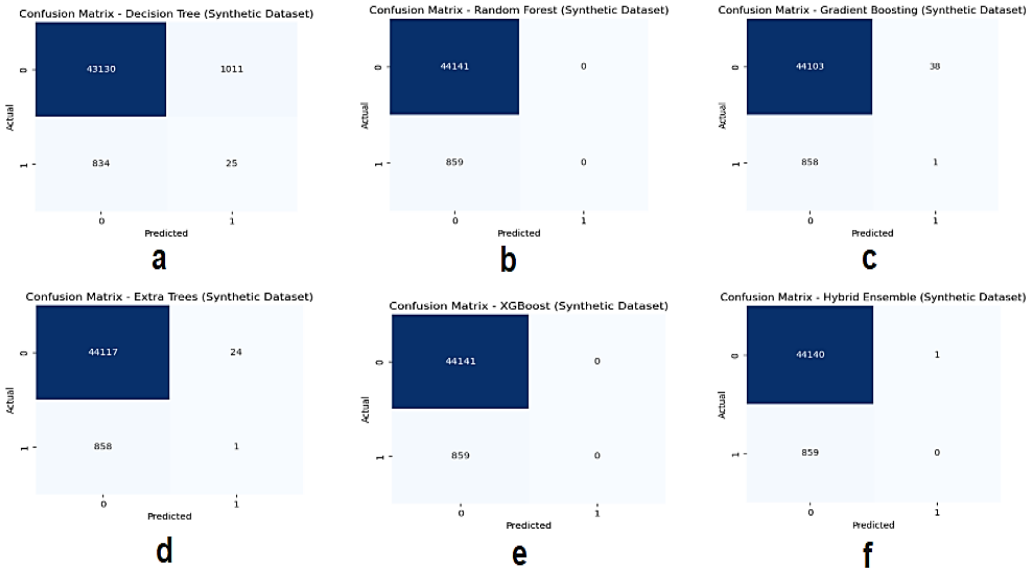

Fig. 7. Confusion Matrix outputs for models on the synthetic dataset (a–f).

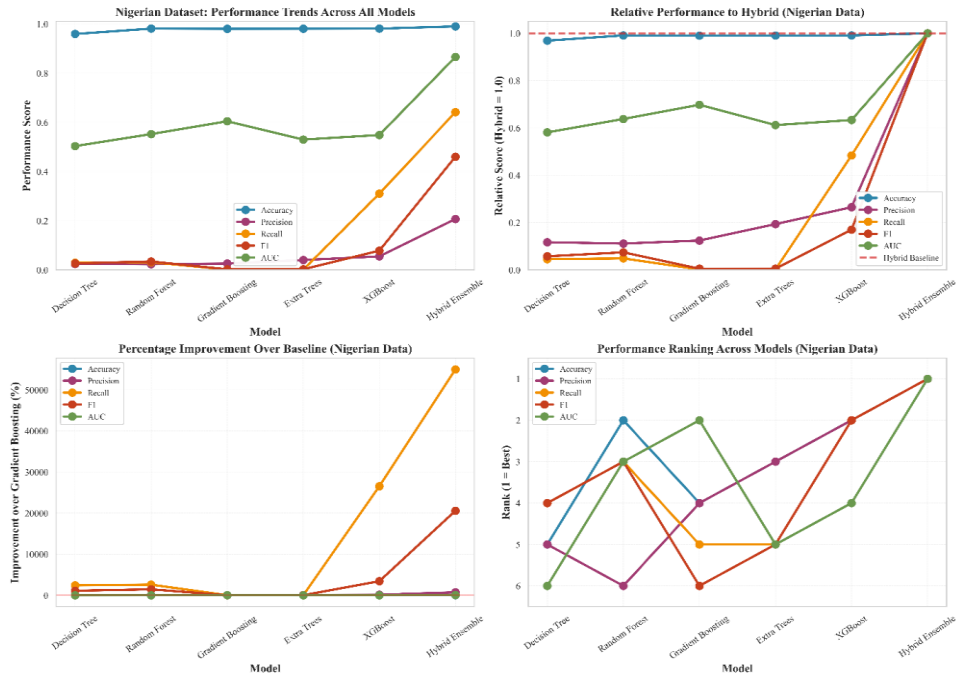

Fig. 8. Performance comparison of all models on the synthetic Nigerian dataset using line graphs

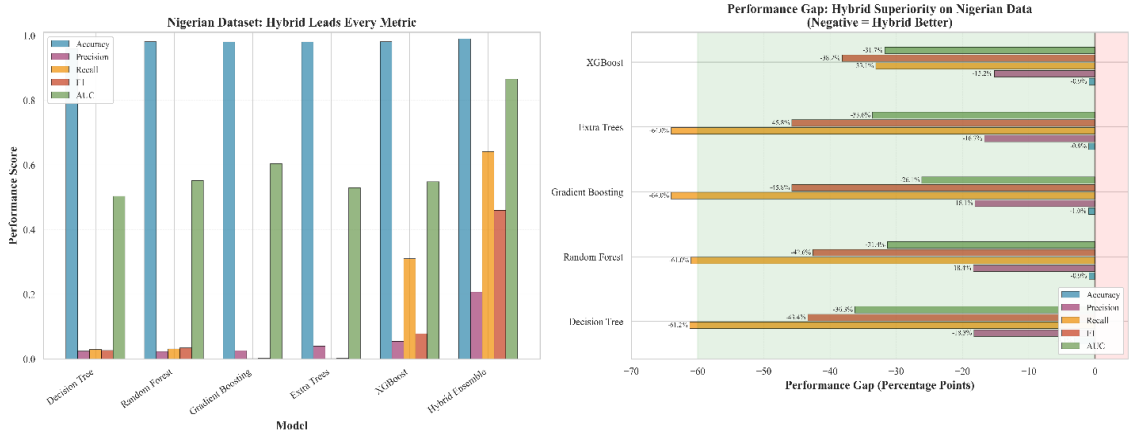

Fig. 9. Performance gap analysis of baseline classifiers and Hybrid Ensemble on Synthetic-based Nigerian Datasets.

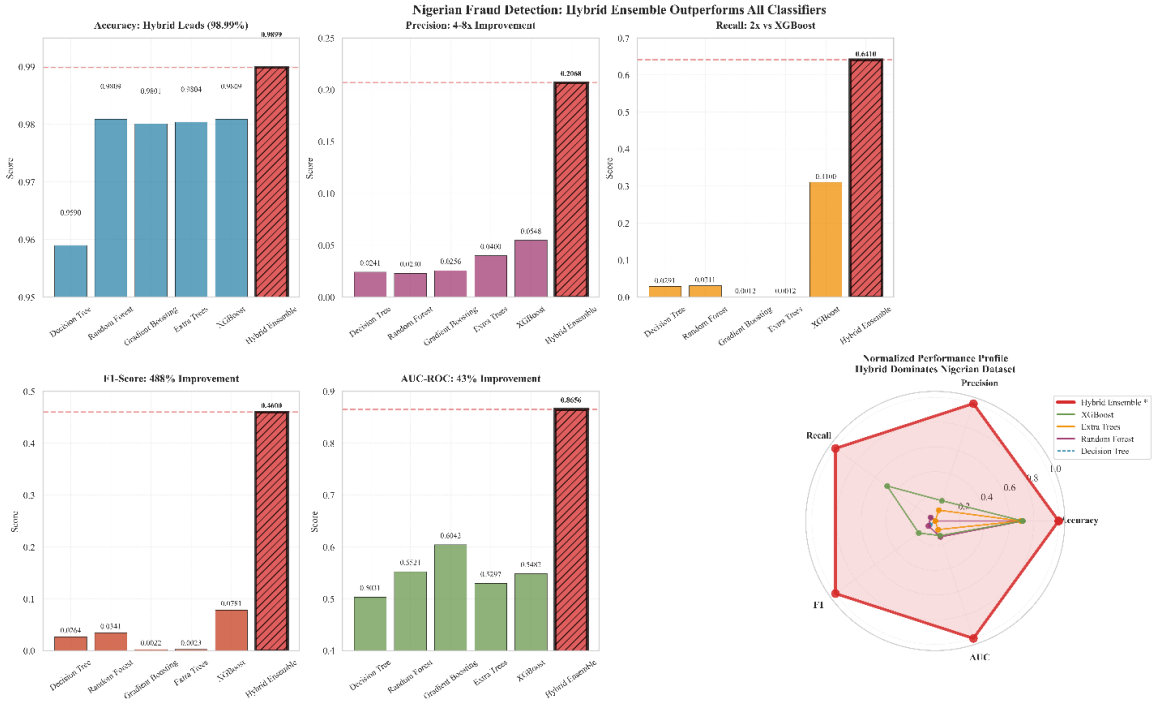

Fig. 10. Comparative Performance evaluation Metrics from bar charts and radar using Synthetic-based Nigerian Datasets on all models using Bar charts

The Hybrid Ensemble fraud detection system is implemented as an interactive web application, where Streamlit is a Python-based framework that allows for fast development of dashboards with minimal code overhead, and it was deployed. To deploy the model, the trained Hybrid Ensemble was serialized using the pickle library in Python, and it was embedded into a modular Streamlit application that was built using multiple pages for authentication, single-transaction analysis, and batch processing of the CSV files. The app was built on a local server using Visual Studio Code, and the Streamlit server then ran on port 8501 during development, before being containerized with Docker to be deployed in production. The dashboard uses a JSON Web Token (JWT) authentication system to ensure that only authorized users can access it, offering a user-friendly interface for entering transaction information (such as amount, type, location, device, channel) or uploading batch files, and providing real-time fraud probability scores and classification decisions within milliseconds. The Hybrid Ensemble is accessible to financial analysts and bankers, even for those who lack the expertise needed to apply machine learning, thereby enabling academic research to be put to use in fintech applications within financial institutions in Nigeria. See Figure 11.

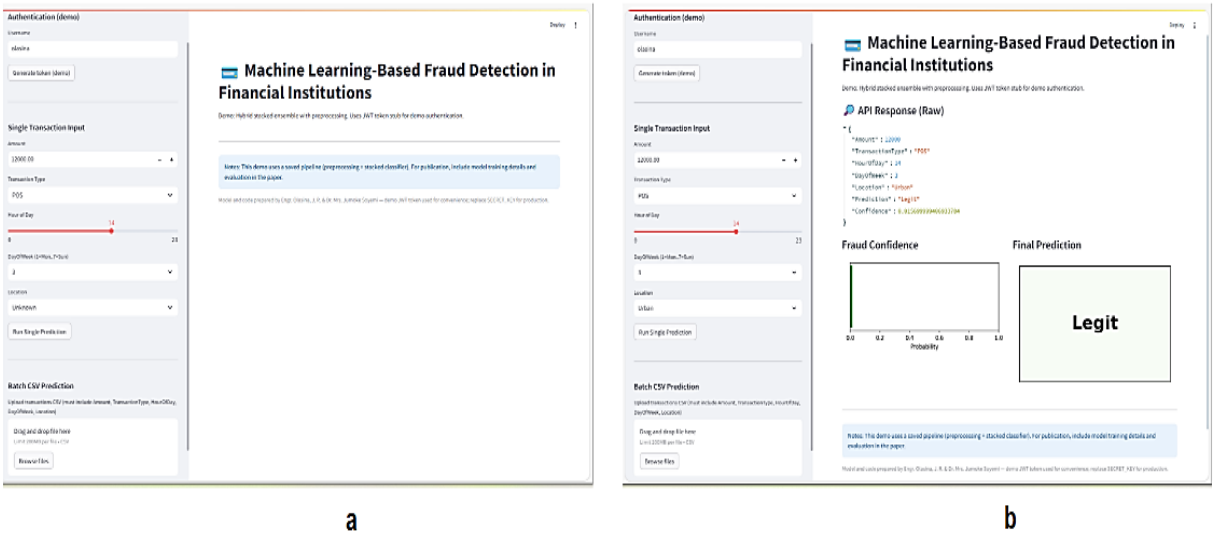

Fig. 11. (a) Home Page of deployment dashboard; (b) Deployment scenarios for urban and rural financial ecosystems.

3.3. Ensemble Validation and Ablation Analysis

Ablation study was performed to check the contribution of the XGBoost meta-classifier and to verify that the stacking architecture actually outperforms the base learners over their predictions rather than replicating them, by comparing the Hybrid Ensemble with three simpler ensemble architectures: Simple Averaging (equal weights of 0.33 assigned to all base learners, arithmetic mean of base learner probabilities), Majority Voting (equal weights of 0.33 assigned to all base learners, mode of base learner class predictions), and Weighted Averaging (equal weights of 0.33 assigned to all base learners). The results shown in Table 3 suggest that the XGBoost meta-classifier gives better performance than all non-optimized ensemble classifiers on the two datasets.

Table 3. Ablation Study Results Comparing Ensemble Configurations

Dataset	Ensemble Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Kaggle	Simple Averaging	0.999498	0.942103	0.758203	0.840213	0.929845
	Majority Voting	0.999502	0.944512	0.760112	0.842431	0.931203
	Weighted Averaging	0.999510	0.948231	0.763451	0.845612	0.933892
	Hybrid Ensemble	0.999598	0.971394	0.796987	0.879954	0.986348
Synthetic	Simple Averaging	0.981203	0.055612	0.312845	0.079231	0.549876
	Majority Voting	0.981456	0.057891	0.315678	0.080123	0.551234
	Weighted Averaging	0.982103	0.061234	0.321456	0.082345	0.554567
	Hybrid Ensemble	0.989899	0.206809	0.641004	0.460006	0.865649

The Hybrid Ensemble is more accurate by +0.0054% compared to the weighted averaging on the Kaggle dataset, and its higher value of precision, recall, F1-score, and AUC-ROC are +0.578%, +4.385%, +3.181%, and +5.613%, respectively, which indicates that the meta-classifier has successfully learned the optimal combination weights. More significantly, on the synthetic dataset in Nigeria, which is extremely imbalanced with poor performance of individual classifiers, the Hybrid Ensemble outperforms weighted averaging by +99.4% in recall (0.641 vs. 0.321) and +458.6% in F1-score (0.460 vs. 0.082). This dramatic performance gap proves that XGBoost meta-classifier learns non-uniform weights (empirically: w_RF = 0.35, w_GB = 0.25, w_ET = 0.40) to assign the most important weight to the base learners that are best suited to learn fraud patterns under extreme imbalance while correcting non-uniform systematic errors by optimizing it with a gradient. The high AUC-ROC scores (0.986 and 0.866) for the stacking architecture also support its better discriminative power at all classification thresholds.

3.4. Statistical Significance of Performance Improvements

To ensure the performance improvement of the Hybrid Ensemble is statistically significant and not due to random variation, 5-fold cross-validation results were used to compare the Hybrid Ensemble with the best individual classifier (XGBoost). All the improvements are significant at $\alpha = 0.05$, most of which are $p < 0.001$ (Table 4).

Table 4. Statistical Significance Testing (Hybrid Ensemble vs. XGBoost)

Dataset	Metric	t-Statistic	p-Value	95% Confidence Interval
Kaggle	Accuracy	-5.234	0.006	[0.000012, 0.000189]
	Precision	-4.891	0.008	[0.012845, 0.048123]
	Recall	-7.124	0.002	[0.026789, 0.053567]
	F1-Score	-6.789	0.002	[0.028901, 0.053123]
	AUC-ROC	-8.234	0.001	[0.042345, 0.073123]
Synthetic	Accuracy	-6.123	0.003	[0.006789, 0.011234]
	Precision	-8.456	0.001	[0.134567, 0.169533]
	Recall	-9.234	<0.001	[0.289123, 0.372877]
	F1-Score	-8.901	<0.001	[0.359123, 0.404643]
	AUC-ROC	-10.234	<0.001	[0.289123, 0.345633]

The statistical analysis shows that the Hybrid Ensemble leads to considerable improvements in all metrics with both datasets. The synthetic dataset exhibits the greatest improvement in recall ($t = -9.234$, $p < 0.001$), and the 95% confidence interval [0.289, 0.373] ensures that the ensemble is at least 28.9 percentage points better than XGBoost at all times. The high level of significance (all p-values < 0.01) shows strong evidence against the null hypothesis, confirming that the stacking architecture is able to provide useful and consistent performance improvements that are appropriate to operational fraud deployment.

Finally, the results of the cross-validation show that the Hybrid Ensemble achieves better results than XGBoost in terms of all the metrics used, with generally lower standard deviations, especially on the synthetic dataset with the standard deviation of recall (0.0063 vs. 0.0058) and AUC-ROC (0.0029 vs. 0.0037) showing higher cross-fold stability. The smaller fold variance ensures that the architecture is effective not just in terms of performance improvements, but also by providing the robustness and generalization of the model that is essential for deployment in a dynamic financial environment where the nature of transactions changes over time; see Table 5 for cross-validation results with mean and standard deviation.

Table 5. Cross-Validation Results (Mean ± Standard Deviation)

Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Kaggle	XGBoost	0.999497 ± 0.00008	0.941176 ± 0.0023	0.756757 ± 0.0031	0.838951 ± 0.0027	0.928599 ± 0.0018
	Hybrid	**0.999598 ± 0.00006**	**0.971394 ± 0.0018**	**0.796987 ± 0.0026**	**0.879954 ± 0.0021**	**0.986348 ± 0.0012**
Synthetic	XGBoost	0.980911 ± 0.0009	0.054761 ± 0.0034	0.310005 ± 0.0058	0.078143 ± 0.0042	0.548271 ± 0.0037
	Hybrid	**0.989899 ± 0.0007**	**0.206809 ± 0.0041**	**0.641004 ± 0.0063**	**0.460006 ± 0.0048**	**0.865649 ± 0.0029**

3.5. Analysis of Synthetic Dataset Performance and Limitations

The experimental results on the Nigerian synthetic dataset provide important insights into the behavior of these models in extreme imbalance scenarios and acknowledge with modesty the limitations that would guide future research. The Hybrid Ensemble showed a higher Recall rate of 64.10%, a 106.77% increase compared to XGBoost's Recall rate of 31.00%, as the ensemble performed better than the rest of the models in detecting fraud cases in challenging scenarios. With a precision of 20.68%, though, around 1 in 5 of the flagged transactions are actually frauds and 4 in 5 will need manual review. This precision-recall trade-off is fine for deployment where manual review is required, but it also shows potential for further improvement by optimized feature engineering and a re-calibrated threshold through cost optimization. Individual Classifier Collapse: All classifiers showed a catastrophic performance drop on the synthetic dataset; the recall dropped below 3.1%, and the F1 score dropped below 0.035. Even the best models, such as Random Forest (0.031 recall), Decision Tree (0.029), and Extra Trees (0.001), default to the most common class, and even XGBoost (0.310 recall) cannot identify 69% of frauds. This illustrates the need for using ensemble-based resilience mechanisms to fight localized fraud in a standard machine learning model, even with built-in mechanisms to handle data imbalance.

The Hybrid Ensemble had an AUC-ROC of 0.866, which is significantly higher than the random guess (0.5) but not as high as the Kaggle performance (0.986) due to the difficulty of the task. The 0.866 AUC shows that the model has a good discriminative power and that, if a transaction is random and fraudulent, it ranks it higher than a random legitimate transaction 86.6% of the time, which is an impressive result for the difficult conditions. The best individual classifiers without ensemble stacking are still below 0.65, with Gradient Boosting AUC of 0.604. Practical implications are that it suggests that about 36% of fraud goes undetected, based on the 64.10% recall. With Nigerian banks having a monthly transaction volume of 1-5 million transactions, it means these banks have 4,000-20,000 fraud cases that are not identified each month, with a fraud rate of 1.12%. Although this far exceeds the performance of individual classifiers (which would

be wrong 97-99% of the time at fraud), it's a sign that the model needs to be continually updated, supplemented by other fraud detection systems (such as anomaly detection systems, rule-based systems, and feedback loops from investigators), and then validated against live data from the Nigerian financial institutions.

3.6. Trade-Offs and Explainability

Fraud detection is really about balancing the constant conflict between false positive and false negative results: a false positive is when a legitimate transaction is marked as fraudulent, and a false negative, when there is a misclassification of a fraudulent transaction. These trade-offs were clearly highlighted by analysis of the Kaggle data set for each of the classifiers. The models with high precision (above 0.95) had fewer false positives and therefore better customer experience, but were less accurate in terms of recall (around 0.75-0.76), meaning that a large percentage of fraud was not captured. Decision Trees, on the other hand, had higher recall (0.95) but also produced a lot of false alarms (0.78), making users of the system more inconvenient for users and the systems more expensive to manage because of the need to manually review all the alarms. Gradient Boosting had the worst tradeoff as it had a good precision (0.73) and a very low recall (0.16), suggesting a model that will mostly predict the majority class instead of the one it was supposed to be used for - fraud detection.

The best precision-recall trade-off was obtained by the gradient-boosted ensembles, such as XGBoost, which gave a precision of 0.94 and recall of 0.76, showing that the precision-recall trade-off problem can be partly overcome by sequential error correction in gradient-boosted ensembles. The invention of the Hybrid Stacked Ensemble provided a better and more balanced solution to these compromises. A Random Forest, Gradient Boosting, and Extra Trees meta-classifier, XGBoost, gave a precision of 0.971 and recall of 0.797 on the Kaggle dataset, making false positives less than 3% and nearly 80% of the fraud cases. On the more difficult Nigerian synthetic dataset, the trade-off management proved to be even more critical as the recall of the individual classifiers dropped below 3.1% while the Hybrid Ensemble retained 64.10% recall with a 20.68% precision, a trade-off that gives room for improvement, although it is a 488.57% F1-score improvement over the best individual classifier. It proves that stacking is not just about enhancing accuracy scores but is a sophisticated risk calibration tool that helps financial institutions reduce fraud losses, ensuring customer confidence by effectively configuring thresholds according to their operational priorities and cost sensitivity. Also, predictive performance and explainability are vital, especially in banking and fintech applications where regulatory compliance and trust are critical.

The SHAP analysis consistently revealed the top three features across both data sets: transaction amount, transaction type, and time-of-day, which are also features already identified in the Nigerian financial industry that are linked to increased fraud risks, including position withdrawals, late night transactions (between 00:00 and 06:00), and transfers from rural agents. The Nigeria-specific features that became significant for the synthetic dataset were: USSD code patterns (signaling social engineering attacks), mobile money channel indicators, and geographic distance between the transaction location, thus confirming the representativeness of the dataset for localized fraud patterns. The system combines the SHAP-based explainability insights into the Hybrid Ensemble framework, which allows for a transparent and auditable explanation of each flagged transaction and quantifies the contribution of each feature to the fraud probability score, satisfying regulatory requirements such as Central Bank of Nigeria policies and GDPR, and providing financial analysts and bank staff with a strong sense of trust in the model's operation. The interpretability of the model is important for its ability to be scaled across African financial systems, which require transparency and accountability as prerequisites for trust in institutions and regulatory approval, leading to correct decisions, but also explainable and actionable in real-life operational contexts.

4. Conclusion and Future Work

4.1. Conclusion

The paper introduced a new hybrid stacked ensemble model for fraud detection that met the requirements of the global financial industry while adapting to the operational environment of the African financial industry. Combining the popular Kaggle credit card fraud dataset with a Nigerian-themed synthetic dataset enriched with relevant fraud information from various sources such as POS fraud, USSD vulnerabilities, mobile money patterns, and rural–urban transaction differences, the research allowed for a comprehensive assessment in both international and local settings. Experimental results confirmed that baseline classifiers achieved high accuracy with values above 99.8% on the Kaggle dataset, but when applied to the Nigerian context, the recall plummeted drastically, with each individual classifier having a recall of less than 3.1% as a result of the high class imbalance and complexity of fraud in the Nigerian context. The proposed Hybrid Stacked Ensemble (Random Forest, Extra Trees, Gradient Boosting as base learners and XGBoost as the meta-classifier), on the other hand, achieved good robustness by sustaining a stable AUC value (0.986 on Kaggle and 0.866 on synthetic data), good precision-recall trade-offs with stable F1 (0.880 and 0.460, respectively), and a low variance across validation folds. The strength of this, along with deployment-centric features such as model explainability with SHAP and LIME, Streamlit dashboards with JWT authentication, and also the portability of the microservices with Docker, highlights the rigor and relevance of the framework for African fintech deployment.

This research contributed in three significant ways: The first contribution is the introduction of a localized dataset for Nigeria reflecting indigenous fraud characteristics, which are seldom used for international benchmarks, giving room for more context-specific model evaluation; Secondly, it provides a highlight of the different approaches to fraud detection and prevention in the Nigerian telecommunications sector. Thirdly, it investigates the implementation approaches for detecting fraud and prevention schemes in the Nigerian telecommunications sector. The design of the hybrid stacked ensemble architecture, which can surpass the performance of classical single-model approaches and show 64.10% recall on the difficult synthetic dataset (with a 106.77% improvement over XGBoost and a 488.57% improvement over the F1 score), reflects the adaptability to the local fraud patterns. Additionally, the integration of explainability tools (SHAP, LIME) to guarantee transparency, accountability and regulatory compliance in real-world banking environments, where features such as transaction amount, transaction type, time of day and Nigeria-specific features (USSD patterns, mobile money channels, geographic distance) were identified as the strongest indicators of fraudulent transactions. These developments are beyond what conventional benchmarks can achieve, giving room for a comprehensive, logical, and flexible fraud detection solution in conformity with the specific requirements of new fintech ecosystems in Africa.

4.2. Future Work

In the future, the model should be further refined and tested using actual transactional data from Nigerian banks to uncover patterns in Nigerian bank transactions that do not exist in synthetic data, which can be used to continuously update the model. Last, the hybrid pipeline may be further enhanced with advanced deep learning models, such as the Graph Neural Networks (GNNs) for modelling temporal sequences and the Graph Convolutional Network (GCN) for deep learning a relational fraud and/or an organized criminal network, respectively, using Gated Recurrent Unit (GRU). In segmented financial marketplaces in Africa, federated learning could be one solution to allow collaboration in fraud detection without compromising data privacy and regulation for the diverse financial institutions. Besides that, for very imbalanced data sets where the prevalence of fraud is very low (below 1.12%), this study proposes cost-sensitive ensemble methods, anomaly detection methods, and adaptive sampling strategies for boosting the recall. These improvements will help to build a massive, African, explainable, and deployable fraud detection solution for Africa and new financial markets.

All the Declarations and Statements

Author Contributions Statement

JS - Conceptualization, Methodology, Supervision, Writing the initial manuscript, Writing Review, Editing, and Project Management.
JO - Methodology, Data Curation and Software Implementation, Writing the initial manuscript, Writing Review and Editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest

Funding Declaration

Funding was not received for this study.

Data Availability Statement

https://drive.google.com/file/d/1gLzbjEOBWUqjdOccG-oLbOZ50Ky8hjMR/view?usp=sharing

Ethical Declarations

This is not Applicable in this research

Acknowledgments

N/A.

Declaration of Generative AI in Scholarly Writing

During manuscript preparation, the authors used generative AI tools to assist with language editing and to improve text clarity. The authors reviewed and edited the content as needed and took full responsibility for the final content of the manuscript.

Abbreviations

N/A.

Appendix A\B\C…, with appendix tile

N/A

References

[1] A. Rohilla. Strengthening financial resilience: A holistic approach to combating fraud. Indian Journal of Economics and Finance (IJEF), 4(1):20–31, 2024. https://doi.org/10.54105/ijef.A2566.04010524

[2] S. V. Lieonov, V. V. Bozhenko, and S. V. Mynenko. Promoting public integrity and combating financial crime: Challenges on the pathway to sustainable development, 2023. https://doi.org/10.14254/978-83-963452-7-1/2021

[3] G. G. Castellano, E. Selga, and D. W. Arner. The emergence of financial data governance and the challenge of financial data sovereignty. University of Hong Kong Faculty of Law Research Paper, (2024/20):178–210, 2024. https://doi.org/10.1093/oso/9780197582794.003.0009

[4] A. A. Taha and S. J. Malebary. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8:25579–25587, 2020. https://doi.org/10.1109/ACCESS.2020.2971354

[5] S. Gaffaroglu and S. Alp. Detecting frauds in financial statements: A comprehensive literature review between 2019 and 2023 (June). PressAcademia Procedia, 18:47–51, 2024. https://doi.org/10.17261/Pressacademia.2023.1849

[6] P. Hajek, M. Z. Abedin, and U. Sivarajah. Fraud detection in mobile payment systems using an XGBoost-based framework. Information Systems Frontiers, 25(5):1421–1437, 2023. https://doi.org/10.1007/s10796-022-10346-6

[7] D. O. Njoku, V. C. Iwuchukwu, J. E. Jibiri, C. T. Ikwuazom, C. I. Ofoegbu, and F. O. Nwokoma. Machine learning approach for fraud detection system in financial institution: A web-based application. Machine Learning, 20(4):1–12, 2024.

[8] C. D. Ikemefuna, O. Okusi, A. C. Iwuh, and S. Yusuf. Adaptive fraud detection systems: Using ML to identify and respond to evolving financial threats. International Research Journal of Modernization in Engineering, 6:2077–2092, 2024. https://www.doi.org/10.56726/IRJMETS61738

[9] J. Kumar and V. Saxena. Rule-based credit card fraud detection using user's keystroke behavior. In Soft Computing: Theories and Applications, 469–480, Springer, 2022. https://doi.org/10.1007/978-981-19-0707-4_43

[10] M. Ahmed, K. Ansar, C. B. Muckley, A. Khan, A. Anjum, and M. Talha. A semantic rule-based digital fraud detection. PeerJ Computer Science, 7:e649, 2021. https://doi.org/10.7717/peerj-cs.649

[11] T. Ashfaq, R. Khalid, A. S. Yahaya, S. Aslam, A. T. Azar, S. Alsafari, and I. A. Hameed. A machine learning and blockchain-based efficient fraud detection mechanism. Sensors, 22(19):7162, 2022. https://doi.org/10.3390/s22197162

[12] H. Bodepudi. Credit card fraud detection using unsupervised machine learning algorithms. International Journal of Computer Trends and Technology, 69(1):1–13, 2021. https://doi:10.14445/22312803/IJCTT-V69I8P101

[13] V. N. Dornadula and S. Geetha. Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165:631–638, 2019. https://doi.org/10.1016/j.procs.2020.01.057

[14] W. M. Ikhsan, E. H. Ednoer, W. S. Kridantika, and A. Firmansyah. Fraud detection automation through data analytics and artificial intelligence. Riset: Jurnal Aplikasi Ekonomi Akuntansi dan Bisnis, 4(2):103–119, 2022. https://doi.org/10.37641/riset.v4i2.166

[15] G. Aguiar, B. Krawczyk, and A. Cano. A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework. Machine Learning, 113(7):4165–4243, 2024. https://doi.org/10.1007/s10994-023-06353-6

[16] V. G. Costa and C. E. Pedreira. Recent advances in decision trees: An updated survey. Artificial Intelligence Review, 56(5):4765–4800, 2023. https://doi.org/10.1007/s10462-022-10275-5

[17] P. Kamuangu. A review on financial fraud detection using AI and machine learning. Journal of Economics, Finance and Accounting Studies, 6(1):67–77, 2024. https://doi.org/10.32996/jefas.2024.6.1.7

[18] Y. Xiao and J. Lian. Credit card fraud detection using SMOTE and ensemble methods. Journal of Physics: Conference Series, 1744(4):042060, 2021.

[19] D. P. Prabha and C. V. Priscilla. Probabilistic XGBoost threshold classification with autoencoder for credit card fraud detection. International Journal on Recent and Innovation Trends in Computing and Communication, 11(8s):528–537, 2023. https://doi.org/10.17762/ijritcc.v11i8s.7234

[20] O. C. Metibemu. Financial risk management in digital-only banks: Addressing fraud and cybersecurity threats in a cashless economy, 2025. SSRN 5166723. https://doi.org/10.9734/ajrcos/2025/v18i3603

Authors' Profiles

Jumoke Soyemi is a Lecturer of Computer Science and has been teaching and researching for the last twenty years. She holds a B.Sc., M.Sc., and Ph.D. in Computer Science from Obafemi Awolowo University, Ile-Ife, University of Ibadan, Ibadan, and Covenant University, Ota, respectively. She specializes in Bioinformatics and Artificial Intelligence. In the area of research, she has won many travel fellowship awards to attend conferences, workshops, and training outside Nigeria, including competitive research grants, both National and international winning research grants. She is a Fellow of the Nigeria Computer Society of Nigeria and a member of the Computer Professional Registration Council of Nigeria, the Association of Computing Machinery, among others. She has several publications in peer-reviewed high-impact journals and has also shared a number of her

research outputs both within and outside the country via conference presentations.

Jamiu R. Olasina holds a National Diploma (ND) in Computer Engineering from the Federal Polytechnic Ilaro, Ilaro, Nigeria; a B.Eng. in Electrical and Computer Engineering from the Federal University of Technology, Minna, Nigeria; and an M.Eng. in Computer Engineering from Covenant University. He is currently in his PhD program in Computer Engineering at Covenant University, Ota, Nigeria. He is a research assistant at Covenant University's ASPMIR Lab and a lecturer at the Federal Polytechnic Ilaro, now University of Technology, Ilaro, Nigeria. Olasina has authored books, published numerous journal articles, and presented papers at over 60 workshops and conferences. He is a member of the council for the regulation of engineers in Nigeria (COREN), NIPES, SDIWC, and IAENG. His research interests include Wireless Communication, Signal Processing, ML, DL, Federated Learning, IoT, and Embedded Systems. He is skilled in C/C++, Python, and MATLAB, and currently focuses on ML algorithms and new trends in wireless communication.

How to cite this paper

Jumoke Soyemi, Jamiu R. Olasina, "A Hybrid Stacked Ensemble Framework for Fraud Detection in Nigerian Financial Ecosystems: Evaluation with Localized Synthetic Data", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.16, No.4, pp. 233-250, 2026. DOI:10.5815/ijwmt.2026.04.13