Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-1 | Jan 2026
Original Researcher Article

Detecting And Preventing Financial Fraud in Banks Using AI and Big Data Analytics

Lokendra Patel [1], Dr.Priyanka Gupta [2]*
[1]Research Scholar - Nims School Of Law Nims University Rajasthan Jaipur
Email Id - Adv.Lokendrapatel@Gmail.Com
[2]Associate Professor, Nims School Of Law Nims University Rajasthan Jaipur
Email Id - Priyanka.Gupta@Nimsuniversity.Org

ABSTRACT
The rapid growth of digital banking and online financial services has significantly increased the risk and complexity of financial fraud, demanding intelligent and scalable detection mechanisms. Traditional rule-based systems are often inadequate due to high false-positive rates and limited adaptability to evolving fraud patterns. To address these challenges, this study proposes an AI- and Big Data–driven fraud detection framework that integrates machine learning and deep learning techniques for accurate and real-time fraud identification. The proposed methodology employs XGBoost and Long Short-Term Memory (LSTM) models, along with a novel hybrid LSTM–XGBoost architecture, to capture both transactional patterns and temporal behavioral characteristics from large-scale banking transaction data. Extensive experiments conducted on a real-world benchmark dataset demonstrate the effectiveness of the proposed approach. The hybrid model achieves superior performance with an accuracy of 0.989, precision of 0.907, recall of 0.946, F1-score of 0.926, and AUC of 0.987, while also significantly reducing the false positive rate to 0.021. Furthermore, scalability analysis confirms its suitability for big data environments with efficient training and low inference latency. Overall, the results indicate that the proposed framework offers a robust, accurate, and scalable solution for fraud detection in modern banking systems.
Keywords: Financial Fraud Detection; Artificial Intelligence; Big Data Analytics; Hybrid LSTM–XGBoost; Banking Security

1. .INTRODUCTION

The rapid digitization of banking services has fundamentally transformed the way financial institutions operate, offering unprecedented convenience, speed, and accessibility to customers. Online banking platforms, mobile payment systems, real-time fund transfers, and digital wallets have become integral to modern financial ecosystems. However, this digital transformation has also expanded the attack surface for financial fraud, making banks increasingly vulnerable to sophisticated and large-scale fraudulent activities. Financial fraud—ranging from credit card fraud and identity theft to money laundering and cyber-enabled scams—poses a serious threat to the stability, profitability, and reputation of banking institutions. As fraudsters continuously evolve their techniques, traditional rule-based and manual fraud detection systems have become inadequate in identifying complex, fast-moving, and hidden fraud patterns.

In the contemporary banking environment, financial fraud is no longer limited to isolated or easily detectable incidents. Instead, it often manifests as highly organized, data-driven operations that exploit system vulnerabilities, customer behavior, and transaction complexity. Conventional fraud detection methods typically rely on predefined rules, thresholds, and historical patterns. While such approaches can detect known fraud scenarios, they struggle to adapt to new and emerging fraud strategies. Moreover, rule-based systems often generate high false-positive rates, leading to unnecessary transaction declines, increased operational costs, and diminished customer trust. These limitations highlight the urgent need for intelligent, adaptive, and scalable fraud detection frameworks capable of processing vast volumes of heterogeneous financial data in real time. Figure 1 shows the advantages of detecting fraud in banks.

Figure 1: Advantages of Detecting fraud in banks

Artificial Intelligence (AI) and Big Data analytics have emerged as transformative technologies in addressing the growing challenges of financial fraud in banks. AI techniques, including machine learning and deep learning, enable systems to learn from historical and real-time data, identify subtle anomalies, and continuously improve detection accuracy without explicit programming. By modeling complex, non-linear relationships within transaction data, AI-driven systems can uncover hidden fraud patterns that are often imperceptible to traditional methods. At the same time, Big Data analytics provides the computational and architectural foundation required to handle massive volumes of structured and unstructured data generated by banking operations, such as transaction logs, customer profiles, behavioral data, device information, and external threat intelligence.

The integration of AI with Big Data analytics allows banks to move from reactive fraud detection to proactive and preventive fraud management. Instead of merely

identifying fraud after it occurs, intelligent systems can predict suspicious behavior, assess risk levels in real time, and trigger timely interventions. This capability is particularly critical in an era where digital transactions occur at high velocity and across multiple channels. Big Data platforms facilitate the ingestion, storage, and processing of high-velocity data streams, while AI algorithms analyze these streams to detect deviations from normal behavior patterns. As a result, banks can monitor transactions continuously, reduce response time, and minimize financial losses.

Another significant advantage of AI-driven fraud detection lies in its ability to enhance decision-making while improving customer experience. Advanced analytics can distinguish between genuine customer behavior and fraudulent activity more accurately, thereby reducing false alarms and unnecessary transaction blocks. This balance is crucial for maintaining customer satisfaction in highly competitive banking markets. Furthermore, AI systems can incorporate contextual and behavioral features—such as spending habits, transaction frequency, geographical location, and device usage—to build comprehensive risk profiles for individual customers. Such personalized analysis strengthens fraud prevention mechanisms while ensuring seamless service delivery.

Despite its potential, the adoption of AI and Big Data analytics in fraud detection also presents notable challenges. Issues related to data quality, data privacy, regulatory compliance, model transparency, and ethical considerations must be carefully addressed. Banking data is highly sensitive, and improper handling can lead to serious legal and reputational consequences. Additionally, the "black-box" nature of some AI models raises concerns regarding interpretability and accountability in critical financial decision-making. Therefore, developing robust, explainable, and compliant AI-based fraud detection systems remains an active area of research and innovation.

In this context, detecting and preventing financial fraud in banks using AI and Big Data analytics has become a vital research and practical domain. The convergence of intelligent algorithms with large-scale data processing technologies offers a powerful solution to combat evolving fraud threats in the digital banking landscape. By leveraging predictive analytics, anomaly detection, and real-time monitoring, banks can strengthen their fraud defense mechanisms, protect customer assets, and enhance overall financial security. This topic is therefore of significant importance for researchers, practitioners, and policymakers seeking to build resilient, secure, and future-ready banking systems in an increasingly data-driven world. The research objectives for this study are as follow:

- To develop an intelligent fraud detection framework using AI and Big Data analytics for banking systems.
- To analyze large-scale transaction data through effective preprocessing and feature engineering techniques.
- To evaluate and compare machine learning, deep learning, and hybrid models for fraud detection accuracy.
- To examine the scalability and computational efficiency of the proposed framework for real-time banking applications.

2. LITERATURE REVIEW

The reviewed literature consistently demonstrates that the rapid digitization of banking ecosystems has fundamentally reshaped the nature of financial fraud, necessitating a transition from traditional rule-based detection systems to AI- and Big Data–driven frameworks. Islam et al. (2025) [1] demonstrate that rapid digitalization has rendered traditional rule-based fraud detection ineffective, advocating AI-driven hybrid ensemble models for banking fraud detection. Their large-scale real-world evaluation shows significant improvements in detection accuracy, false-positive reduction, and real-time processing. Moreover, the integration of explainable AI and federated learning addresses transparency and regulatory concerns. Building on this, Iseal et al. (2025) [2] examine the broader role of AI and Big Data analytics in financial services, highlighting their effectiveness in real-time fraud detection and risk assessment. However, they caution that data privacy, regulatory compliance, and algorithmic bias remain key challenges. Similarly, Berrada et al. (2025) [3] review AI and Big Data applications in commercial banking, identifying fraud detection as a high-impact yet underutilized area. Their study emphasizes the importance of data preprocessing, dataset size, and scalable machine learning models for robust fraud analytics.

From an implementation perspective, Ghimire et al. (2025) [4] analyze AI–Big Data–based fraud detection systems deployed in major U.S. banks, demonstrating improved real-time fraud identification and loss reduction. They also discuss emerging technologies such as blockchain and biometrics while noting concerns around data security and system adaptability. At the model level, Sujana et al. (2025) [5] propose a hybrid CNN–LSTM framework that captures both spatial and temporal transaction patterns, outperforming traditional machine learning techniques with lower false positives. Extending this technical focus, Emran et al. (2024) [6] conduct a PRISMA-based systematic review, confirming the effectiveness of deep learning, ensemble models, and NLP in detecting complex fraud patterns. In parallel, Angela et al. (2024) [7] explore fraud prevention in FinTech environments, highlighting behavioral biometrics and blockchain integration as effective complements to Big Data–driven machine learning models.

Within a broader financial security context, Ahmad et al. (2023) [8] highlight the synergy between AI and Big Data in strengthening fraud analytics and cybersecurity through real-time anomaly detection and predictive risk management. In the Indian banking scenario, Eni et al. (2023) [9] discuss how AI and Big Data adoption supports advanced fraud detection and real-time analytics, while also raising concerns about privacy, skills, and regulatory compliance. Focusing on infrastructure, Sekar et al.

(2023) [10] propose a cloud-based real-time fraud detection framework using optimized feature selection and machine learning classifiers to improve scalability and responsiveness. Complementing this, Venigandla et al. (2022) [11] show that integrating robotic process automation with AI-driven predictive analytics enhances fraud detection efficiency and operational speed. Furthermore, Pattabhi et al. (2022) [12] emphasize the role of AI-enabled decision support systems in fraud detection and regulatory compliance, advocating explainable AI and strong governance frameworks. Finally, Hassan et al. (2021) [13] provide foundational evidence that hybrid AI and Big Data approaches significantly improve anomaly detection accuracy and computational efficiency, forming the basis for modern real-time fraud detection systems in banking.

Although prior studies demonstrate the effectiveness of AI and Big Data analytics in fraud detection, gaps remain in developing unified frameworks that balance accuracy, false positive reduction, and scalability. Limited work comprehensively evaluates hybrid models under large-scale, real-time banking conditions while incorporating behavioural feature engineering and class imbalance handling. Moreover, practical deployment aspects such as inference efficiency and customer impact are often underexplored.

3. RESEARCH METHODOLOGY

- Dataset Description

The Bank Transaction Dataset for Fraud Detection from Kaggle [1] is a comprehensive transactional dataset created to support research in financial fraud detection and anomaly analysis. It contains simulated bank transaction records capturing a mix of legitimate and potentially fraudulent activities, with every transaction labeled to indicate whether it is fraudulent, enabling supervised learning. The dataset includes key transactional and temporal attributes such as transaction date and time, customer identifiers, transaction amount, and indicators of previous transaction history. This structure allows researchers to perform feature engineering, time-based analysis, and machine learning model development to detect irregular patterns that signify fraud. The dataset's realistic distribution, class imbalance (with far fewer fraudulent cases than legitimate ones), and diversity of features make it ideal for evaluating AI and Big Data–oriented fraud detection frameworks where both accuracy (e.g., recall, precision) and computational efficiency are critical.

Table 1: Dataset Features Table

Feature Name	**Data Type**	**Description**
TX_FRAUD	Categorical (0/1)	Binary label indicating whether the transaction is fraudulent (1) or legitimate (0).
TX_DATETIME	Date/Time	Timestamp when the transaction occurred, useful for temporal pattern analysis.
TX_AMOUNT	Numeric	The monetary value involved in the transaction.
TX_PREV_AMOUNT	Numeric	Amount of the immediately previous transaction by the same customer.
TX_PREV_DATETIME	Date/Time	Timestamp of the previous transaction, enabling gap/time-difference features.
TX_FRAUD_SCENARIO	Categorical	Category indicating the type or scenario of fraud (if tagged).
CustomerID	Categorical	Unique customer identifier for behavior profiling.
Channel	Categorical	Indicates channel/medium used (e.g., ATM, mobile, web).
OtherTransactionFeatures…	Mixed	Additional attributes related to transaction behavior and history.

3.2 Data Preprocessing and Big Data Handling

Given the large-scale and heterogeneous nature of the bank transaction dataset, a structured Big Data preprocessing pipeline is adopted to ensure statistical robustness and computational efficiency. Missing numerical values are handled using statistically informed imputation methods. For a numerical feature x, missing entries are replaced by the median value , computed as

which is preferred over the mean due to its robustness to outliers commonly present in transaction data. Missing categorical values are imputed using the mode or assigned a dedicated "unknown" category to preserve data integrity.

Categorical variables are encoded using **target encoding** and **frequency encoding** to retain predictive information. In target encoding, a categorical feature ccc is transformed into the conditional probability of fraud as

where denotes the fraud label and represents the subset of records with category c. Frequency encoding maps each category to its relative occurrence in the dataset, defined as

where N is the total number of transactions. These encodings are particularly suitable for high-cardinality

banking features such as customer IDs and transaction channels.

To support gradient-based AI models, numerical features are standardized using z-score normalization:

where μ and σ represent the mean and standard deviation of the feature, respectively. This scaling ensures comparable feature ranges and improves convergence stability during training.

Since financial fraud datasets are inherently imbalanced, the **Synthetic Minority Oversampling Technique (SMOTE)** is applied to the training data to augment minority (fraudulent) samples. For a minority class instance , a synthetic sample is generated as

where is one of the k-nearest neighbors of in the minority class. This approach balances class distribution while preserving the underlying data structure.

Finally, the dataset is partitioned into training, validation, and testing sets using **stratified sampling**, ensuring consistent fraud–non-fraud proportions across splits. Formally, for each subset S, the class ratio is maintained as

where D denotes the full dataset. This preprocessing framework enables scalable, unbiased, and mathematically grounded preparation of banking transaction data for AI- and Big Data–driven fraud detection models.

3.3 Feature Engineering and Selection

Feature engineering is performed to extract meaningful behavioral and transactional patterns that enhance fraud detection accuracy. Temporal features are derived to capture transaction dynamics, including transaction frequency and time gaps. For a user u, transaction frequency within a time window Δt is defined as

where represents the number of transactions in the interval. The inter-transaction time gap is computed as

with unusually short gaps often indicating fraudulent activity.

Aggregated statistical features summarize user-level behavior. The mean transaction amount is calculated as

where denotes transaction amounts. Additional statistics such as variance and maximum values capture spending irregularities. Interaction features are also derived, such as normalized transaction amount

which highlights deviations from typical customer behavior.

Feature selection is applied to reduce dimensionality and improve computational efficiency. Pearson correlation is used to measure the association between features and the fraud label:

Low-correlation features are removed. Additionally, tree-based feature importance is computed using impurity reduction:

Only highly informative features are retained, improving model interpretability, scalability, and fraud detection performance.

3.4 AI Model Development

The proposed fraud detection framework is designed using a **three-model strategy** to effectively capture the complex, non-linear, and temporal characteristics of large-scale banking transaction data. Two standalone models and one hybrid model are employed to ensure robustness, interpretability, and high predictive performance.

Gradient Boosting (XGBoost)

XGBoost is selected due to its strong performance on structured financial data and its inherent interpretability. The model predicts the fraud probability by aggregating the outputs of multiple decision trees:

where denotes the space of regression trees. Training minimizes a regularized objective function:

which controls model complexity and reduces overfitting, making it suitable for high-dimensional fraud data.

Long Short-Term Memory (LSTM)

To capture temporal dependencies in sequential transaction behavior, an LSTM network is employed. The LSTM updates its internal states using gating mechanisms:

enabling the model to learn long-term transaction patterns associated with fraudulent activity.

Hybrid Model (LSTM–XGBoost)

The hybrid model integrates the strengths of both approaches. The LSTM acts as a feature extractor, generating high-level temporal representations , which are then provided as input to the XGBoost classifier:

This hybrid architecture enhances fraud detection accuracy while preserving interpretability.

All models are trained using optimized hyperparameters obtained through k-fold cross-validation:

to improve generalization and prevent overfitting, ensuring suitability for real-world banking fraud detection systems.

3.5 Model Evaluation and Performance Metrics

Model performance is evaluated using fraud-specific metrics that are critical in banking applications. These include accuracy, precision, recall, F1-score, Area Under the ROC Curve (AUC), and false-positive rate. Emphasis is placed on recall and AUC to ensure effective fraud detection while minimizing customer inconvenience caused by false alerts. Comparative analysis is conducted between baseline models and the proposed AI-based framework to validate performance improvements.

Result and Analysis

4.1 Feature Importance Analysis

The XGBoost model computes feature importance based on the cumulative reduction in impurity across decision trees, enabling identification of the most influential predictors of fraudulent behavior. This analysis reveals that transaction amount deviations and temporal behavioural features play a dominant role in distinguishing fraudulent transactions from legitimate ones. Table 2 presents the top ten most influential features identified by the XGBoost model for fraud detection. Behavioural and temporal attributes, such as normalized transaction amount, inter-transaction time gap, and transaction frequency, exhibit the highest importance, indicating that deviations from normal spending patterns and transaction timing are key indicators of fraudulent activity.

Table 2: Top 10 Important Features

Rank

Feature Name

Importance Score

1

Normalized Transaction Amount

0.214

2

Inter-Transaction Time Gap

0.187

3

Transaction Frequency (Δt)

0.162

4

Previous Transaction Amount

0.119

5

Channel Encoding

0.097

6

Customer Spending Variance

0.072

7

Maximum Transaction Amount

0.061

8

Fraud Scenario Type

0.045

9

Time of Day

0.029

10

Transaction Count per User

0.014

The figure 2 shows the relative importance of features obtained from the XGBoost model, quantified using importance scores. The **normalized transaction amount** is the most influential feature with a score of **0.214**, followed by **inter-transaction time gap (0.187)** and **transaction frequency (0.162)**, indicating the strong impact of spending deviations and temporal behavior on fraud detection.

Figure 2: Feature Importance Ranking

The **previous transaction amount** contributes moderately with an importance score of **0.119**, while **channel encoding (0.097)** and **customer spending variance (0.072)** have noticeable influence. Features such as **maximum transaction amount (0.061)** and **fraud scenario type (0.045)** show lower but meaningful contributions, whereas **time of day (0.029)** and **transaction count per user (0.014)** have relatively minimal impact on the model's predictions.

4.2 Model Performance Comparison

This section evaluates and compares the classification performance of traditional, machine learning, and deep learning models using standard evaluation metrics. The concern is to identify the most reliable model for accurate and robust fraud detection. Table 4 presents a comparative analysis of model performance across accuracy, precision, recall, F1-score, and AUC. Logistic Regression shows limited discriminative ability, while XGBoost and LSTM achieve substantial improvements. The Hybrid LSTM–XGBoost model delivers the best overall performance, demonstrating superior balance between detection accuracy and reliability.

Table 4: Performance Comparison of Models

Model

Accuracy

Precision

Recall

F1-Score

AUC

Logistic Regression (Baseline)

0.941

0.312

0.684

0.428

0.812

XGBoost

0.982

0.863

0.891

0.877

0.964

LSTM

0.976

0.821

0.913

0.865

0.971

Hybrid LSTM–XGBoost

0.989

0.907

0.946

0.926

0.987

The XGBoost model in figure 3 demonstrates strong performance across all evaluation metrics. It achieves an **accuracy of 0.982**, indicating high overall correctness. The **precision of 0.863** reflects effective reduction of false positives, while the **recall of 0.891** shows strong capability in identifying fraudulent transactions. The balanced **F1-score of 0.877** confirms consistent precision–recall trade-off. Additionally, the **AUC value of 0.964** signifies excellent discrimination between fraudulent and legitimate cases.

Figure 3: Performance Comparison of XGBoost model

The LSTM model exhibits strong predictive performance across all evaluation metrics shown in figure 4. It achieves an **accuracy of 0.976**, reflecting high overall classification correctness. The **precision of 0.821** indicates effective control of false positives, while the **recall of 0.913** highlights its strong ability to identify fraudulent transactions. The **F1-score of 0.865** demonstrates a good balance between precision and recall. Additionally, the **AUC value of 0.971** confirms excellent discriminative capability between fraud and legitimate transactions.

Figure 4: Performance comparison of LSTM Model

The Hybrid XGBoost–LSTM model in figure 5 demonstrates superior performance across all evaluation metrics. It attains an **accuracy of 0.989**, indicating near-perfect classification capability. The **precision of 0.907** reflects a substantial reduction in false positives, while the **recall of 0.946** shows excellent detection of fraudulent transactions. The **F1-score of 0.926** confirms a strong balance between precision and recall. Furthermore, the **AUC value of 0.987** highlights outstanding discriminatory power between fraudulent and legitimate transactions.

Figure 5: Performance comparison of Hybrid (XGBoost-LSTM Model)

The figure 6 shows clear numerical differences in performance among the four models. Logistic Regression achieves an accuracy of **0.941**, but its precision (**0.312**) and F1-score (**0.428**) are low, indicating poor fraud identification despite acceptable recall (**0.684**) and AUC (**0.812**). XGBoost improves performance with **0.982 accuracy**, **0.863 precision**, **0.891 recall**, **0.877 F1-score**, and **0.964 AUC**. The LSTM model records **0.976 accuracy**, **0.821 precision**, **0.913 recall**, **0.865 F1-score**, and **0.971 AUC**. The Hybrid LSTM–XGBoost model outperforms all others, achieving **0.989 accuracy**, **0.907 precision**, **0.946 recall**, **0.926 F1-score**, and **0.987 AUC**, confirming its superior fraud detection capability.

Figure 6: Performance Comparison on different models

7.4 ROC Curve Analysis

The ROC–AUC graph in figure 7 illustrates the comparative classification performance of four fraud detection models: Logistic Regression, XGBoost, LSTM, and the Hybrid LSTM–XGBoost model. The curve plots the True Positive Rate (Recall) against the False Positive Rate. Logistic Regression shows the weakest performance with an AUC of 0.812, remaining closer to the random classifier. XGBoost and LSTM demonstrate strong discrimination with AUC values of 0.964 and 0.971, respectively. The Hybrid LSTM–XGBoost model consistently dominates the ROC space, achieving the highest AUC of 0.987, indicating excellent ability to distinguish fraudulent transactions from legitimate ones, especially at lower false positive rates.

Figure 7: ROC Curves for Fraud Detection Models

7.5 False Positive Rate Analysis

The graph illustrates the False Positive Rate (FPR) comparison among four fraud detection models: Logistic Regression, XGBoost, LSTM, and the Hybrid LSTM–XGBoost model. Logistic Regression exhibits the highest FPR at approximately 0.083, indicating a greater number of false alerts.

Figure 8: Comparison of FPR Value

XGBoost significantly reduces the FPR to about 0.029, while the LSTM model shows a slightly higher FPR of around 0.034. The Hybrid LSTM–XGBoost model achieves the lowest FPR at nearly 0.021, demonstrating superior control over false positives. This substantial reduction in false alerts highlights the hybrid model's suitability for deployment in operational banking environments, as it minimizes unnecessary customer inconvenience while maintaining effective fraud detection performance.

7.6 Scalability and Big Data Efficiency

The proposed framework demonstrates strong scalability when evaluated on large-scale transaction datasets. It efficiently processes high volumes of data with minimal increase in computational overhead, ensuring stable performance. This makes the model well suited for real-

time deployment in big data–driven financial environments.

The **training time graph** in figure 9 illustrates the computational cost of model learning under large-scale data conditions. XGBoost exhibits the lowest training time of **42 minutes**, indicating high efficiency during model fitting. The Hybrid LSTM–XGBoost model requires **55 minutes**, reflecting additional complexity due to deep feature learning combined with ensemble optimization. The LSTM model records the highest training time at **61 minutes**, highlighting the greater computational demand of sequential neural network training on large transaction datasets.

Figure 9: Training time and inference time comparison

The **inference time per transaction graph in figure 9** demonstrates the real-time scalability of the models. XGBoost achieves the fastest inference at **3.8 ms per transaction**, making it highly suitable for low-latency environments. The Hybrid LSTM–XGBoost model maintains a moderate inference time of **4.2 ms**, balancing accuracy and efficiency. In contrast, the LSTM model shows the highest inference time of **6.5 ms**, indicating increased processing overhead during real-time fraud detection.

Conclusion

This study presented an AI- and Big Data–driven framework for detecting and preventing financial fraud in banking systems, addressing the limitations of traditional rule-based approaches. Experimental results demonstrate that advanced models significantly enhance fraud detection accuracy and reliability. The baseline Logistic Regression achieved an accuracy of 0.941 but suffered from low precision (0.312) and a high false positive rate (0.083), limiting its practical applicability. In contrast, XGBoost and LSTM delivered strong improvements, achieving accuracies of 0.982 and 0.976, with AUC values of 0.964 and 0.971, respectively. The proposed Hybrid LSTM–XGBoost model outperformed all others, achieving the highest accuracy of 0.989, precision of 0.907, recall of 0.946, F1-score of 0.926, and AUC of 0.987, while reducing the false positive rate to 0.021. Scalability analysis further showed that the hybrid model maintained efficient inference time (4.2 ms per transaction), confirming its suitability for real-time, large-scale banking environments

REFERENCES

N\A

.