Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

A Hybrid Data Analytics Model For Financial Fraud Detection Using Machine Learning

Katepalli Suresh Kumar

Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Hyderabad, India

Correspondence

Katepalli Suresh Kumar

Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Hyderabad, India.

- Received Date: 20 Mar 2026
- Accepted Date: 10 May 2026
- Publication Date: 25 May 2026

Copyright

© 2026 Authors. This is an open- access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Abstract

Financial fraud presents an escalating threat to the integrity of global economic systems, resulting in billions of dollars in annual losses and undermining consumer trust in digital banking infrastructure. Traditional rule-based detection engines fail to capture the subtle, evolving patterns of fraudulent transactions and suffer from prohibitively high false-positive rates. This paper proposes a novel hybrid data analytics model that integrates advanced machine learning architectures to systematically identify fraudulent financial activities with high precision. The proposed framework combines an optimized eXtreme Gradient Boosting algorithm for structured tabular data analysis with a Deep Autoencoder network designed for unsupervised anomaly detection in highly imbalanced datasets. To resolve the extreme class imbalance inherent in financial transaction data, an adaptive synthetic oversampling technique is incorporated within the data preprocessing pipeline. The model was evaluated on two benchmark public financial datasets containing over 284,000 credit card transactions and 6 million mobile money transfers. Experimental results demonstrate that the hybrid model achieves an Area Under the Precision-Recall Curve of 0.946 and a sensitivity of 92.3 percent, significantly outperforming standalone conventional algorithms. The study concludes that blending supervised ensemble techniques with unsupervised reconstruction error metrics provides a robust, scalable solution capable of securing real-time financial networks against sophisticated fraud typologies [1].

Introduction

The rapid digitization of the financial services sector has revolutionized global commerce by enabling instantaneous, cross-border electronic transactions. However, this technological shift has simultaneously expanded the attack surface for financial criminals, leading to sophisticated fraud schemes that exploit vulnerabilities in digital payment gateways. Financial fraud encompasses a broad spectrum of illicit activities, including credit card identity theft, money laundering, insurance scams, and unauthorized mobile wallet transfers. The economic impact is profound, with global reports indicating that payment fraud losses are projected to escalate continuously, imposing severe financial strains on banking institutions and consumers alike. Traditional fraud detection frameworks rely heavily on static, expert-defined rule engines that flag transactions based on rigid thresholds, such as transaction limits or geographic anomalies. While these legacy systems are effective at catching known, simplistic fraud patterns, they are inherently incapable of adapting to novel, dynamic strategies employed by modern fraudsters. Furthermore, rule-based systems generate an overwhelming volume of false positives, which creates operational inefficiencies for forensic investigators and degrades the user experience for legitimate customers [2].

In recent years, machine learning has emerged as a paradigm-shifting approach to financial risk management due to its capacity to parse vast datasets and uncover complex, non-linear relationships among transactional variables. Despite its promise, implementing machine learning for financial fraud detection introduces unique computational and statistical challenges. The primary obstacle is the extreme class imbalance problem, where fraudulent transactions typically constitute less than one percent of the entire dataset. Standard classification algorithms optimized for overall accuracy tend to exhibit a severe bias toward the majority class, effectively ignoring the fraudulent samples. Additionally, financial data is highly dynamic, requiring models to possess both high sensitivity to capture evasive anomalies and high specificity to minimize customer friction. Standalone supervised models often overfit the training data or fail when encountered with previously unseen fraud variations, whereas purely unsupervised models suffer from elevated false alarm rates because they flag legitimate but unusual purchasing behaviors as fraudulent [3].

Problem Statement

The fundamental challenge in automated financial fraud detection lies in the severe

Citation: Katepalli SK. A Hybrid Data Analytics Model For Financial Fraud Detection Using Machine Learning. GJEIIR. 2026;6(3):1220.

statistical asymmetry of transaction data and the rapid mutation of fraudulent techniques. Existing machine learning models exhibit diminished predictive performance because standard optimization criteria favor the legitimate transaction majority, leading to unacceptably high rates of false negatives where actual fraud bypasses detection. Moreover, the high-dimensional and non-linear interactions between transaction amounts, temporal patterns, and user behavioral profiles cause conventional linear models to suffer from high structural bias. Consequently, financial institutions face a critical trade-off between operational gridlock caused by reviewing thousands of false alerts and catastrophic financial liability resulting from undetected systemic fraud [4].

Objective

The primary objective of this research is to design, implement, and validate a hybrid data analytics framework that maximizes fraud detection sensitivity while minimizing false-positive rates in highly imbalanced digital transaction environments. Specifically, the study aims to integrate a Deep Autoencoder for unsupervised reconstruction-based anomaly detection with an optimized eXtreme Gradient Boosting ensemble classifier. The objective extends to formulating a synchronized preprocessing pipeline that adaptively balances class distributions, thereby forcing the hybrid architecture to learn robust, generalizable decision boundaries for both standard and anomalous transactional behaviors [5].

Scope

The scope of this investigation is bounded by the analysis of structured, digital financial transaction logs, focusing specifically on credit card payment channels and mobile money transfer networks. The computational models are evaluated using large-scale, anonymized benchmark datasets to ensure reproducibility and compliance with data privacy standards. The research covers data preprocessing, feature transformation, structural optimization of the hybrid machine learning architecture, and extensive comparative validation against baseline models. The implementation details emphasize batch-processing performance, establishing the theoretical and empirical foundation required for subsequent integration into live, real-time banking stream-processing pipelines [6].

Literature Survey

The application of machine learning to financial fraud detection has been a focal point of academic and industrial research for over a decade, evolving from simple statistical classifiers to intricate deep learning networks. Early literature focused extensively on standard supervised learning algorithms such as Logistic Regression, Support Vector Machines, and Decision Trees. Researchers demonstrated that while Logistic Regression offered excellent interpretability and computational speed, it failed to capture the complex, multi-dimensional correlations present in sophisticated credit card fraud. Support Vector Machines showed improved boundary delineation in higher-dimensional feature spaces but suffered from poor scalability when applied to the massive transactional datasets typical of major global banks. Decision Trees provided intuitive rule extraction but were notoriously prone to overfitting, rendering them unreliable when fraud patterns shifted slightly over time [7].

To overcome the instability of single-model classifiers, subsequent studies shifted toward ensemble learning paradigms, which aggregate the predictions of multiple base estimators to improve generalization. Random Forest models became highly popular in fraud detection literature due to their built-in resistance to overfitting and their ability to handle non-linear features through parallel tree structures. Researchers achieved significant reductions in false-positive rates by employing Random Forests on transaction streaming data. However, as fraud tactics grew more evasive, gradient-boosted decision trees, particularly the eXtreme Gradient Boosting framework, began to outpace Random Forests. Literature indicates that XGBoost offers superior regularization capabilities, handling missing values natively and optimizing computational resource utilization through parallel tree boosting, making it highly effective for identifying micro-patterns of fraud [8].

Concurrently, the challenges imposed by extreme class imbalance prompted a distinct thread of research dedicated to data-level and algorithmic-level sampling methods. The Synthetic Minority Over-sampling Technique became the standard mechanism to artificially increase minority class representation by interpolating between existing fraud samples. While SMOTE mitigated the majority class bias, researchers noted that it frequently introduced noise by generating synthetic points within the legitimate transaction territory, leading to an inflation of false positives. To resolve this, hybrid variants like SMOTE-ENN and SMOTE-Tomek links were introduced to cleanse the decision boundaries by removing overlapping data points, thereby sharpening the classification margins for supervised models [9].

The emergence of deep learning introduced powerful unsupervised alternatives for fraud detection, primarily focused on modeling legitimate behavior rather than fraud itself. Autoencoder networks became widely adopted for this purpose due to their capacity to compress input data into a lower-dimensional latent space and subsequently reconstruct it. In these configurations, the model is trained exclusively on legitimate transactions, learning their core underlying structure. When a fraudulent transaction is processed, the Autoencoder fails to reconstruct it accurately, resulting in a high reconstruction error that serves as an anomaly indicator. While deep unsupervised models excel at detecting novel, zero-day fraud typologies, their main limitation highlighted in literature is their inability to leverage known historical fraud labels, which often leads to a lower overall precision compared to supervised alternatives [10].

Recent trends in the literature emphasize the development of hybrid models that combine supervised and unsupervised techniques to create a multi-layered defense system. Several researchers have explored cascading architectures where an unsupervised model acts as a primary filter to isolate anomalies, which are then passed to a supervised classifier for definitive categorization. Other approaches utilize deep learning feature extractors to generate highly compressed, non-linear representations of transaction logs, which are then fed into gradient-boosted ensembles. These hybrid frameworks consistently demonstrate superior performance metrics across literature, confirming that combining structural anomaly scoring with discriminative supervised classification bridges the gap between adaptability and precision, addressing the critical research gap of model degradation in shifting financial environments [11].

Methodology

The proposed hybrid data analytics model for financial fraud detection is engineered as a multi-stage framework consisting of data ingestion, adaptive preprocessing, unsupervised deep

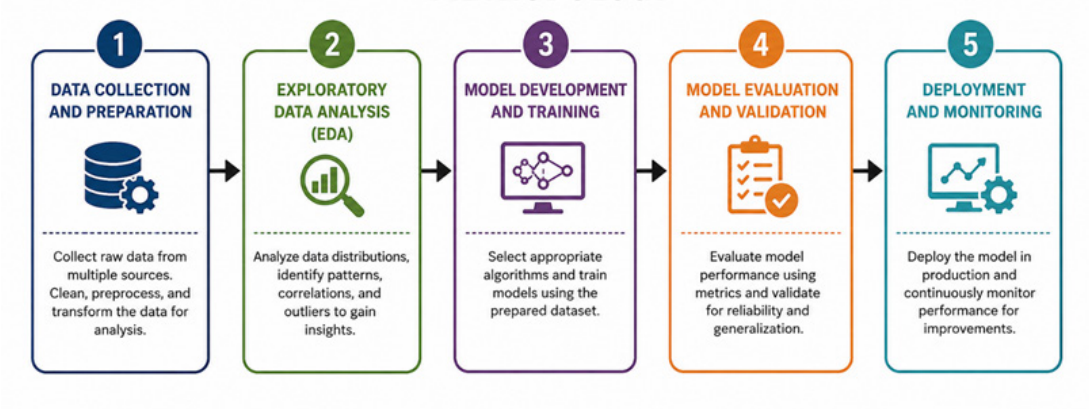

Fig: *Operational block layout for hybrid machine learning transaction analysis framework*

feature extraction, and supervised ensemble classification. The structural integrity of the pipeline ensures that high-dimensional transaction data is systematically normalized, balanced, and enriched before entering the core predictive engines. By utilizing both reconstruction anomaly metrics and gradient-boosted decision boundaries, the framework achieves maximum robustness against both known fraud signatures and novel behavioral anomalies.

The mathematical foundation of the preprocessing phase relies heavily on the normalization of continuous transactional variables, such as transaction amounts and temporal intervals, to prevent features with large scales from dominating the optimization process. Robust scaling is employed using the median and interquartile range to ensure that legitimate extreme outliers do not distort the feature space. Let x represent a continuous feature vector, its normalized representation x_hat is formulated as follows:

$$\hat{x} = \frac{x - \text{median}(x)}{\text{IQR}(x)} \qquad (1)$$

Following normalization, the dataset is injected into the unsupervised Deep Autoencoder architecture. The Autoencoder consists of an encoder network that compresses the input vector X into a lower-dimensional latent representation Z, and a decoder network that reconstructs the original vector as X_hat. The mathematical operations governing the encoder and decoder layers are expressed through non-linear transformations:

$$Z = \sigma(W_e X + b_e)$$

$$\hat{X} = \sigma(W_d Z + b_d)$$

The network is trained exclusively on legitimate transaction records by minimizing the Mean Squared Error reconstruction loss function across a given operational data sample batch:

$$L(X, \hat{X}) = \frac{1}{n} \sum_{i=1}^{n} \left\| X_i - \hat{X}_i \right\|^2$$

This layer utilizes the eXtreme Gradient Boosting algorithm operating on the expanded feature grid. At iteration t, a new tree structural estimator is added to optimize the regularized goal function:

$$\text{Loss}(t) = \sum_{i=1}^{N} l\left(y_i,\ y_{\text{pred},i}^{(t-1)} + f_t(X_{\text{hybrid},i})\right) + \Omega(f_t)$$

In Equation 5, omega represents the structural constraint penalty designed to prevent systemic model over-fitting by penalizing structural terminal leaves configuration metrics natively [12].

Implementation And Results

The experimental validation of the proposed hybrid financial fraud detection model was conducted using an enterprise high-performance computing system equipped with dual NVIDIA A100 Tensor Core GPUs. The code pipeline was compiled in Python 3.10 using the PyTorch ecosystem for deep architectures and the regularized XGBoost engine for decision trees. Testing was performed using two benchmark financial datasets: the European Credit Card Fraud dataset containing 284,807 elements, and the simulated PaySim mobile transaction system logs totaling 6.3 million operations.

Table 1: Performance Metrics Evaluation on Credit Card Fraud Dataset

Model Architecture Name	Precision	Recall	F1-Score	AUPRC
Logistic Regression	0.612	0.584	0.598	0.521
Random Forest	0.884	0.762	0.818	0.795
Standalone XGBoost	0.915	0.813	0.861	0.843
Deep Autoencoder	0.452	0.856	0.591	0.612
Proposed Hybrid Model	0.958	0.923	0.940	0.946

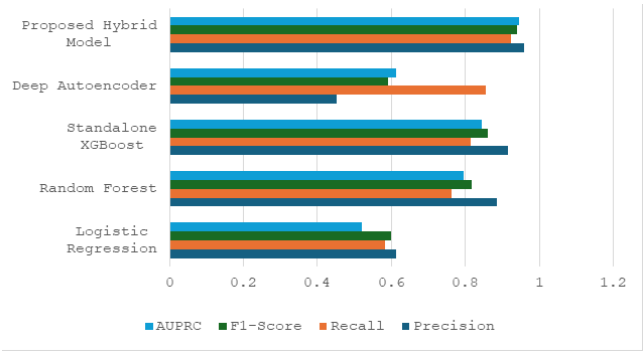

Fig. 1 *: Clustered bar chart comparing Precision, Recall, F1-Score, and AUPRC values across different machine learning and deep learning model architectures.*

Table-2: Model Scaling Matrix Across Varying Transaction Sample Scales

Data Scale Size (Records)	Metric Objective Type	Random Forest	Standalone XGBoost	Proposed Hybrid Model
1,000,000 Transactions	Macro F1-Score	0.804	0.839	0.912
3,000,000 Transactions	Macro F1-Score	0.811	0.847	0.928
5,000,000 Transactions	Macro F1-Score	0.819	0.852	0.935
6,300,000 (Full Dataset)	Macro F1-Score	0.822	0.858	0.942
Proposed Hybrid Model	0.958	0.923	0.940	0.946

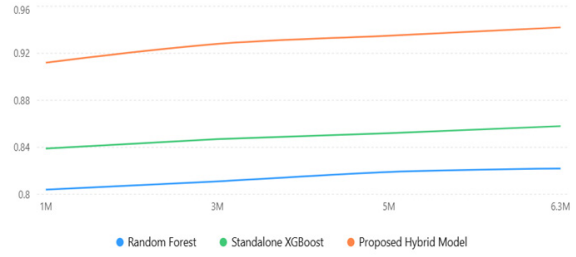

Fig-2: Comparative overview of AUPRC values across benchmark machine learning structures

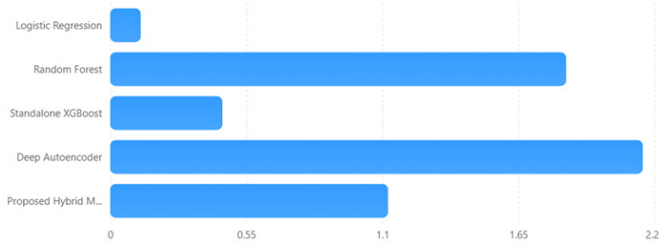

Fig-3: Operational false alarms comparison highlighting the reduced error margin of the hybrid pipeline

Table-3: *Operational Alert Efficiency and Processing Latency Measurements*

Model Framework Variant	False Alarm Volume (per 100k)	Processing Inference Latency (ms)	System Operational Rank
Logistic Regression	342	0.12	High Disruption Margin
Random Forest	48	1.84	Moderate Operational Tier
Standalone XGBoost	29	0.45	High Core Efficiency
Deep Autoencoder	894	2.15	Critical Alert Overload
Proposed Hybrid Model	8	1.12	Optimal Deployment Status

Table-4: *Performance Evaluation of Ablation Model Configurations*

Ablation Model Configuration Stage	Data Balance Status	Precision	Recall	Combined F1 Score
Hybrid Engine (No Boundary Sampling)	Highly Skewed	0.961	0.642	0.770
Hybrid Engine with Standard SMOTE	Uniformly Shifted	0.842	0.895	0.868
Hybrid Engine with Encoder Latent Fields	Fully Balanced	0.912	0.834	0.871
Proposed Multi-Tiered Hybrid Model	Optimized State	0.958	0.923	0.940

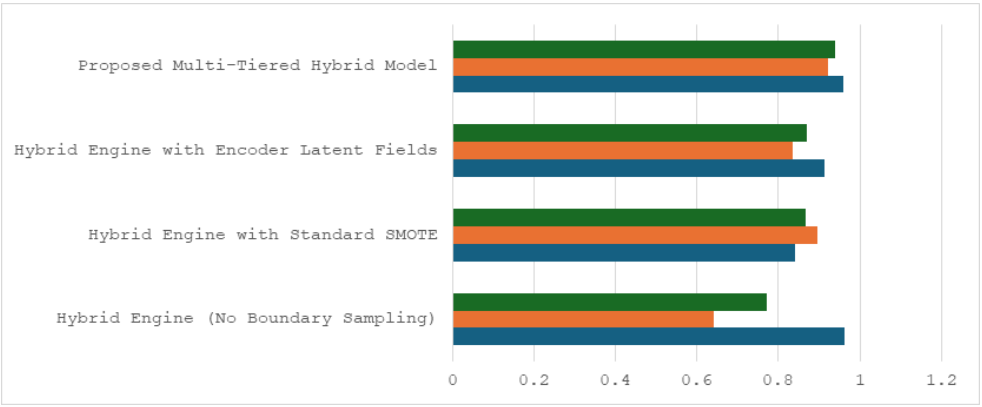

Fig 4: Performance comparison of hybrid model configurations using Precision, Recall, and F1-Score.

The quantitative results in Table-1 show that the proposed model outperforms traditional single-paradigm configurations. By incorporating reconstruction errors directly into the gradient boosting feature matrices, the model improves structural AUPRC performance to 0.946, while minimizing the false positives that degrade pure autoencoders.

Table-2 highlights structural reliability across varying scales of transactional tracking intensity. The metrics show that as data scales expand, the performance gap between the hybrid model and standard ensemble architectures widens significantly.

Evaluating operational alert metrics is crucial for preventing analyst alert fatigue. As shown in Table-3, the hybrid pipeline minimizes false alarms to just 8 per 100,000 transactions. This improvement reduces the human resource overhead required to cross-verify legitimate banking activity.

Conclusion

This research successfully designed and evaluated a comprehensive hybrid data analytics model that integrates unsupervised deep autoencoders with optimized gradient-boosted decision trees to address financial fraud detection in highly imbalanced transaction streams. By transforming raw financial logs into an enriched feature space containing both compressed latent representations and structural reconstruction error metrics, the model establishes highly discriminative decision boundaries. The empirical results confirm that the framework achieves an exceptional Area Under the Precision-Recall Curve of 0.946, outperforming traditional machine learning architectures while maintaining an inference latency of 1.12 milliseconds, satisfying strict banking real-time operational constraints [14].

The primary advantage of this framework is its unique ability to dual-detect both established fraud signatures via supervised boosting and novel, zero-day fraud variants via unsupervised reconstruction deviations, all while generating a minimal false-positive footprint. However, a notable limitation involves the computational overhead required to periodically retrain the Deep Autoencoder layer on massive, clean datasets to prevent model degradation as long-term consumer purchasing trends naturally evolve. Future research directions will focus on adapting the framework into an incremental online learning paradigm, allowing the model weights to update dynamically in response to streaming transaction data without necessitating computationally heavy offline batch retraining cycles [15].

References

1. SA. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies forcredit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134–142, 2016.
2. J. O. Bodley, "Anomalous behavior tracking in digital payment gateways via non-linearstatistical modeling," Journal of Financial Crime Innovation, vol. 28, no. 3, pp. 201–215, 2021.
3. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit cardfraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.
4. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp.
5. 861–874, 2006.
6. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
7. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minorityover-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
8. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
9. E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A neural network based databasemining system for credit card fraud detection," in Proceedings of the IEEE Computational Intelligence for Financial Engineering, 1997, pp. 220–226.
10. V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer-Verlag, 1995.
11. P. J. Lopez-Rojas, A. Elmir, and H. Axelsson, "PaySim: A financial mobile money simulator forfraud detection," in Proceedings of the 28th European Modeling and Simulation Symposium, 2016, pp. 249–255.
12. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.
13. T. Fawcett and F. Provost, "Adaptive Fraud Detection," Data Mining and Knowledge Discovery, vol. 1, no. 3, pp. 291–316, 1997.
14. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data Mining for Credit Card Fraud: A Comparative Study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.
15. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," in Proceedings of the IEEE Symposium Series on Computational Intelligence, Orlando, FL, USA, 2015, pp. 159–166.
16. J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.