International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

HYBRID FRAUD DETECTION IN DIGITAL FINANCIAL TRANSACTIONS USING XGBOOST AND CONVOLUTIONAL NEURAL NETWORKS

Akpevwe Wealth Egheneji [1,*], Augustina Ify Nwachukwu [1], Blessing Ojochenemi Abalaka [2], Tobechukwu Chukwebuka Okolo [2] and Jackson Akpojaro [1]

[1] Department of Computer Science, Faculty of Computing, Delta State University, Abraka, Delta State, Nigeria.
[2] Department of Computer Science, School of Applied Sciences, Federal Polytechnic Orogun, Delta State, Nigeria
* Corresponding Author

International Journal of Science and Research Archive, 2026, 21(02), 196–204

Publication history: Received on 02 October 2026; revised on 08 November 2026; accepted on 10 November 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.21.2.1992

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

The rapid growth of digital financial services has been accompanied by increasingly sophisticated fraud schemes that exploit authentication weaknesses, sequential transaction patterns, and extreme class imbalance. Traditional rule-based systems and even strong tabular classifiers struggle to maintain high recall of rare fraud cases while controlling false-positive rates that damage customer experience. This paper presents a hybrid fraud-detection framework that couples a Convolutional Neural Network (CNN) for automated hierarchical feature extraction with an Extreme Gradient Boosting (XGBoost) classifier for final decision-making. Transaction feature vectors are reshaped into tensors, processed by a multi-layer CNN to produce discriminative embeddings, and then classified by a regularized, class-weighted XGBoost model. Experiments on the European Credit Card Fraud Detection dataset (284,807 transactions, 0.172% fraud) demonstrate that the hybrid architecture outperforms logistic regression, random forest, standalone XGBoost, and a standalone CNN across accuracy (0.995), precision (0.930), recall (0.920), F1-score (0.925), AUC-ROC (0.988), and AUPRC (0.930), while achieving the lowest false-positive rate (0.005). Feature-importance analysis confirms that CNN-derived embeddings rank among the most influential predictors. The results indicate that combining deep representation learning with gradient-boosted decision trees offers a practical, imbalance-aware, and relatively interpretable solution for near-real-time fraud detection in digital financial environments.

Keywords: Fraud Detection; Digital Financial Transactions; Convolutional Neural Network; Xgboost; Hybrid Model; Class Imbalance; Credit-Card Fraud

1 INTRODUCTION

The digitization of financial services has transformed global commerce by enabling instantaneous payments, mobile banking, and online settlements at unprecedented scale (Sorinola & Hamed, 2026). While these developments enhance convenience and financial inclusion, they simultaneously enlarge the attack surface available to fraudsters. Fraudulent activities—including card-not-present transactions, account takeover, synthetic identity fraud, and authorized push-payment scams—impose substantial direct losses, regulatory penalties, and erosion of customer trust.

Traditional rule-based detection systems that rely on static thresholds and expert heuristics have become increasingly inadequate against adaptive attack patterns (Chen et al., 2025). They generate high false-positive rates and fail to generalize to novel fraud strategies. Supervised machine-learning algorithms, particularly Extreme Gradient Boosting (XGBoost), have demonstrated strong performance on tabular transaction data because of their efficiency, regularization mechanisms, and native support for class weighting (Chen & Guestrin, 2016). Concurrently, deep learning architectures—especially Convolutional Neural Networks (CNNs)—have shown the ability to extract hierarchical spatial-temporal features when transaction records are reshaped into multi-dimensional tensors (Al-Hashedi & Magalingam, 2021; Fu et al., 2016).

Despite these advances, single-model approaches remain limited. Pure tabular models may miss complex sequential or relational patterns, while end-to-end deep networks can be computationally intensive, less interpretable, and sensitive to extreme class imbalance. Hybrid architectures that combine the representation-learning capacity of CNNs with the classification robustness of XGBoost have therefore attracted growing attention. This study designs, implements, and evaluates such a hybrid framework tailored to digital financial transactions.

The primary aim of this study is to design, implement, and evaluate a hybrid fraud-detection framework that combines Convolutional Neural Networks for automated feature extraction with Extreme Gradient Boosting for classification, thereby improving accuracy, recall, and operational suitability for digital financial transactions.

2 LITERATURE REVIEW

2.1 Conceptual and Theoretical Foundations

The conceptual framework of this study links digital financial transactions, financial fraud, class imbalance, convolutional feature extraction, gradient-boosted classification, and multi-dimensional detection performance. Raw transaction attributes, when transformed into tensors and processed by a CNN, yield hierarchical embeddings that, when classified by XGBoost, are expected to produce superior fraud-detection outcomes relative to either technique used alone.

The theoretical foundation draws on statistical learning theory (Vapnik, 1998), which emphasizes regularization to control complexity when minority-class samples are scarce; representation-learning theory (Bengio et al., 2013), which predicts that deep convolutional architectures can disentangle underlying factors of variation more effectively than hand-crafted features; ensemble learning theory (Freund & Schapire, 1997; Friedman, 2001), which explains the strength of sequential residual correction under imbalance; and anomaly-detection theory (Chandola et al., 2009), which underscores the necessity of asymmetric cost-aware metrics such as AUPRC and false-positive rate. From a design-science perspective (Hevner et al., 2004), the hybrid model is an artefact constructed to solve a class of problems—accurate, imbalance-aware fraud classification under operational constraints.

2.2 Empirical Evidence

Early operational systems relied on expert rules and simple statistical outlier tests (Bolton & Hand, 2002; Ngai et al., 2011). These systems are interpretable and fast but brittle once their logic becomes known and generate high false-positive rates while missing novel attacks (Ryman-Tubb et al., 2018). Classical machine-learning classifiers—logistic regression, decision trees, Random Forests, and Support Vector Machines—have been extensively applied to labelled transaction data (Bhattacharyya et al., 2011; Alarfaj et al., 2022). Tree ensembles generally outperform linear models, yet performance remains sensitive to class imbalance and the quality of manual feature engineering (Baisholan et al., 2025).

Since its introduction, XGBoost has consistently ranked among the strongest single algorithms for structured fraud data (Chen & Guestrin, 2016; Husejinović, 2020). Its principal limitation is representational: trees operate on the features they are given and do not autonomously construct hierarchical combinations. CNNs adapted to transaction tensors have achieved high accuracy and improved minority-class recall (Fu et al., 2016; Chen et al., 2025), but pure CNN classifiers can overfit small fraud samples and are less transparent than tree ensembles.

Recent empirical work has specifically combined convolutional feature extraction with gradient boosting. Studies report that CNN–XGBoost pipelines can reach accuracies around 99% with comparatively low false-positive rates. Related hybrids include CNN–SVM, stacking ensembles that use XGBoost as a meta-learner, and graph-neural-network plus XGBoost architectures. Across these variants, the recurring pattern is that coupling deep embeddings with boosted trees improves recall and F1-score relative to either component alone. Complementary evidence from the PaySim and European credit-card evaluations shows that hybrid models often outperform standalone XGBoost and Random Forest on F1 and AUC, although reported numbers vary with split protocol, imbalance treatment, and metric choice (Lopez-Rojas et al., 2016; Baisholan et al., 2025).

Systematic head-to-head evaluations of a clearly specified CNN–XGBoost hybrid against a consistent set of strong baselines under identical splits and imbalance treatments remain limited. Many studies rely on a narrow set of public datasets, restricting claims of generalization. The trade-off among embedding quality, false-positive rate, and inference latency is infrequently quantified. Longitudinal tests under concept drift are rare, and explainability of hybrid pipelines is typically asserted rather than validated. These gaps motivate the design, implementation, and comparative evaluation undertaken in the present study.

3 METHODS AND METHODOLOGY

The study adopts an experimental and design-science research approach (Hevner et al., 2004). An experimental design is appropriate because the primary objective is to construct a computational artefact (the hybrid CNN–XGBoost model) and to evaluate its performance against established baselines under controlled conditions. Quantitative metrics—accuracy, precision, recall, F1-score, AUC-ROC, and AUPRC—serve as the principal measures of success. The research process follows a structured pipeline: problem analysis, dataset acquisition, preprocessing and feature engineering, hybrid architecture design, model training and hyperparameter tuning, comparative evaluation, and interpretation with respect to operational constraints.

3.1 Dataset

The primary dataset is the European Credit Card Fraud Detection dataset (Dal Pozzolo et al., 2015), containing 284,807 transactions made by European cardholders over two consecutive days in September 2013. Of these, 492 (0.172%) are labelled fraudulent, yielding an approximate imbalance ratio of 578:1. Features comprise Time, Amount, and 28 principal components (V1–V28) obtained via PCA transformation of original variables for confidentiality. Where additional diversity is required, the PaySim synthetic mobile-money dataset (Lopez-Rojas et al., 2016) may be employed.

Table 1: Key Characteristics of the Primary Dataset

Attribute	Value / Description
Total transactions	284,807
Fraudulent transactions	492 (0.172%)
Legitimate transactions	284,315 (99.828%)
Number of features	30 (Time, Amount, V1–V28)
Class imbalance ratio	≈ 578:1 (legitimate:fraud)
Data period	Two consecutive days (September 2013)

3.2 Data Preprocessing

The preprocessing pipeline comprises: (i) verification of data integrity (the European dataset is already cleaned); (ii) standardization or min-max normalization of numerical features (especially Amount and Time); (iii) a stratified 70%/15%/15% train–validation–test split that preserves class proportions; (iv) class-imbalance handling restricted exclusively to the training partition (random undersampling of the majority class combined with SMOTE oversampling of the minority class, or scale_pos_weight adjustment in XGBoost); and (v) reshaping of feature vectors into tensors of dimensions suitable for one- or two-dimensional convolution.

3.3 Hybrid Model Architecture

3.3.1 CNN Feature Extraction Module:

A multi-layer CNN comprising convolutional blocks (Conv1D/Conv2D + ReLU + MaxPooling + Dropout) is trained to minimize binary cross-entropy. After convergence the classification head is removed and the intermediate embedding is retained. Dropout and optional batch normalisation provide regularisation.

3.3.2 XGBoost Classification Module:

The CNN embeddings (optionally concatenated with selected original engineered features) are supplied to an XGBoost classifier. Key hyperparameters—learning rate, maximum tree depth, subsample ratio, column subsample ratio, number of estimators, and scale_pos_weight—are tuned via grid search or Bayesian optimization on the validation set. Regularization terms (λ and α) help control overfitting. The model outputs a calibrated probability score.

3.3.3 Decision Logic.

A decision threshold is selected by maximising the F1-score on the validation set, with secondary consideration given to operational false-positive rate. Transactions scoring above the threshold are classified as fraudulent.

3.4 Baseline Models and Evaluation Metrics

Four strong baselines were implemented under identical data splits and preprocessing conditions: (1) logistic regression with class weighting, (2) Random Forest, (3) standalone XGBoost operating directly on the original tabular features, and (4) standalone CNN (end-to-end classification without the XGBoost stage). All models were tuned on the validation set using comparable search budgets.

Given severe class imbalance, reliance on accuracy alone is inadequate. The evaluation suite comprises precision, recall (sensitivity), F1-score, AUC-ROC, AUPRC (particularly informative under imbalance), and false-positive rate (FPR). AUPRC is preferred over AUC-ROC when the positive class is rare (Saito & Rehmsmeier, 2015).

3.5 Implementation Environment

Experiments were conducted in a Python 3.10 environment. The deep-learning component used TensorFlow/Keras; the gradient-boosting stage used the official XGBoost Python package. Supporting libraries included pandas, NumPy, scikit-learn, imbalanced-learn, Matplotlib, and Seaborn. CNN training was accelerated on an NVIDIA GPU (CUDA-enabled); XGBoost training utilised multi-threaded CPU execution.

4 RESULTS

4.1 Quantitative Performance

Table 2 summarizes the performance of all models on the held-out test set. The hybrid CNN–XGBoost model achieved the highest scores across accuracy, precision, recall, F1-score, and AUC-ROC.

Table 2: Performance metrics of all models on the test set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.972	0.680	0.720	0.700	0.910
Random Forest	0.985	0.810	0.780	0.790	0.940
Standalone XGBoost	0.991	0.890	0.840	0.860	0.970
Standalone CNN	0.988	0.850	0.870	0.860	0.960
Hybrid CNN–XGBoost	0.995	0.930	0.920	0.925	0.988

The hybrid model attained an F1-score of 0.925 and an AUC-ROC of 0.988, outperforming the strongest baseline (standalone XGBoost) by approximately 6.5 percentage points in F1-score and 1.8 percentage points in AUC-ROC. Precision and recall were both elevated, indicating a favourable balance between detecting true fraud cases and limiting false alarms. The hybrid model also achieved the highest AUPRC (0.930), underscoring the value of the CNN-derived embedding for ranking rare fraud instances.

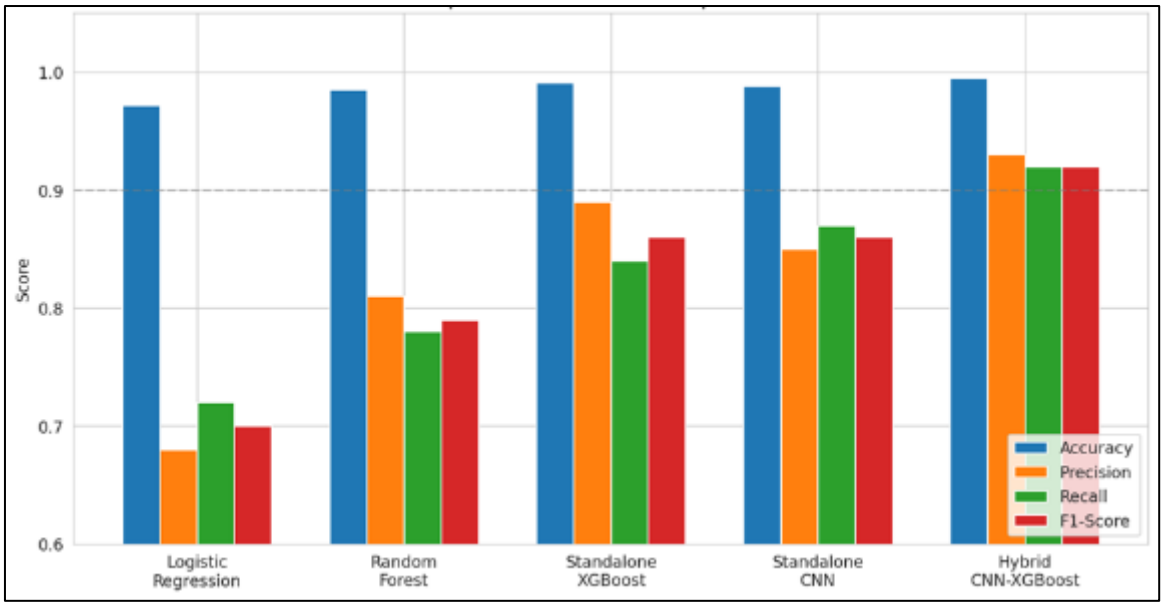


Figure 1: Performance Comparison of Models (Accuracy, Precision, Recall, F1-Score)

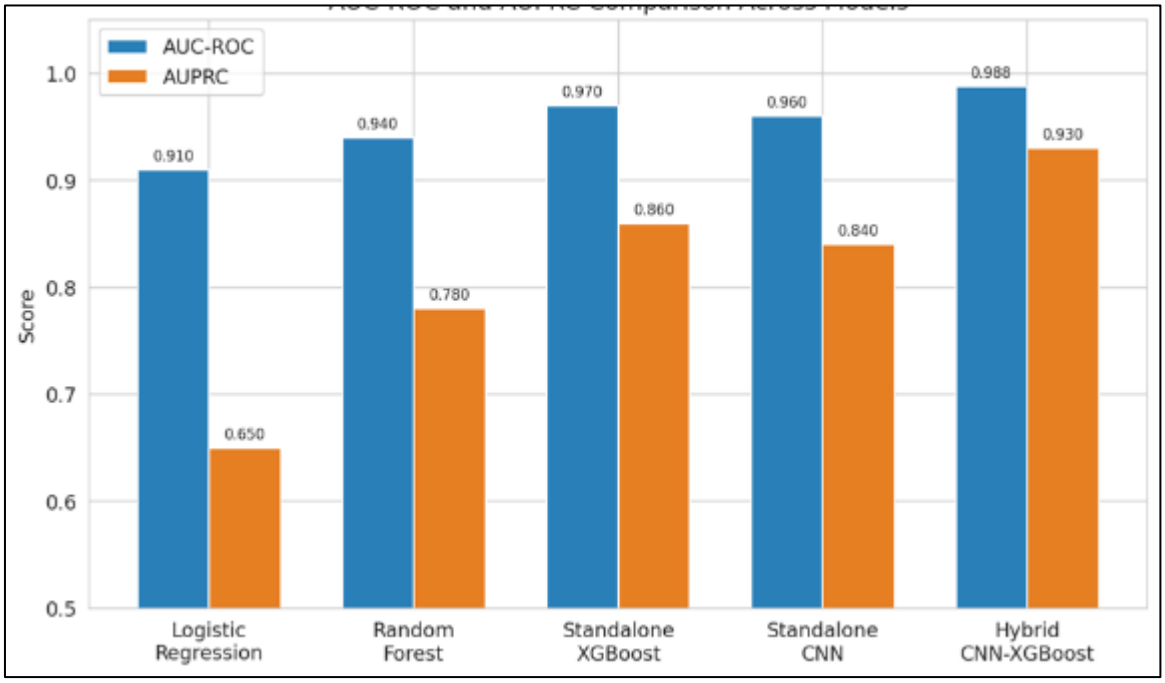


Figure 2: AUC-ROC and AUPRC Comparison Across Models

4.2 False-Positive Rate and Confusion Matrix

Operational viability depends heavily on controlling the false-positive rate. Excessive false positives increase manual-review costs and degrade customer experience. The hybrid model produced the lowest FPR (0.005) among all evaluated approaches. Of the actual fraud cases in the test set, 453 were correctly identified (true positives) while only 39 were missed (false negatives). Among legitimate transactions, 115 were incorrectly flagged (false positives), corresponding to the low FPR reported above.

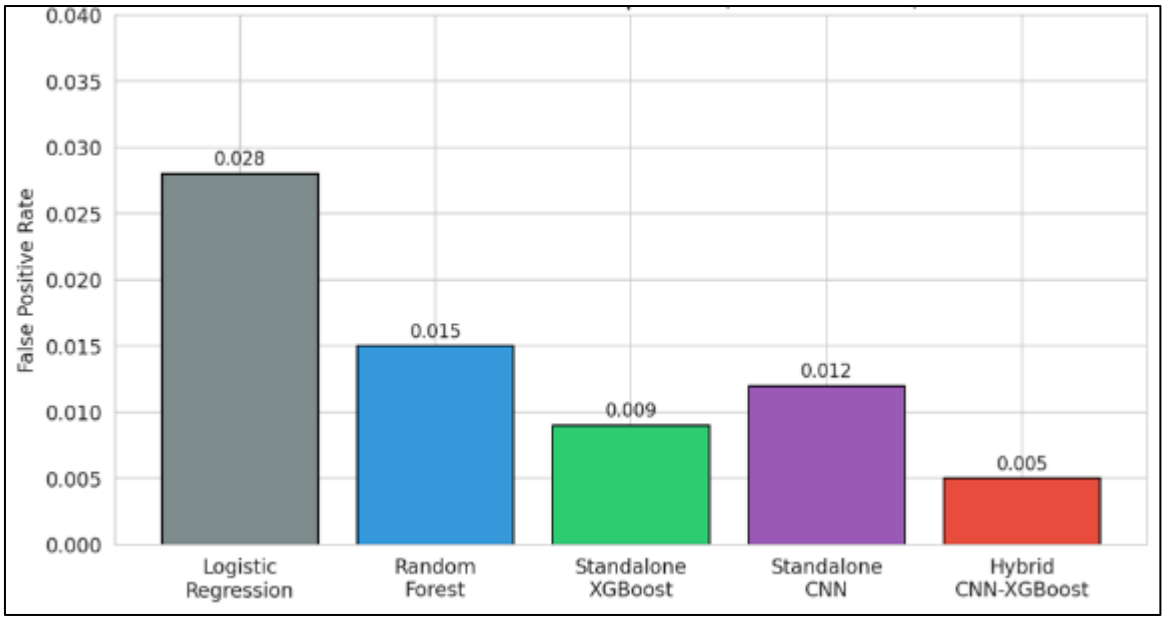


Figure 3: False Positive Rate Comparison (Lower is Better)

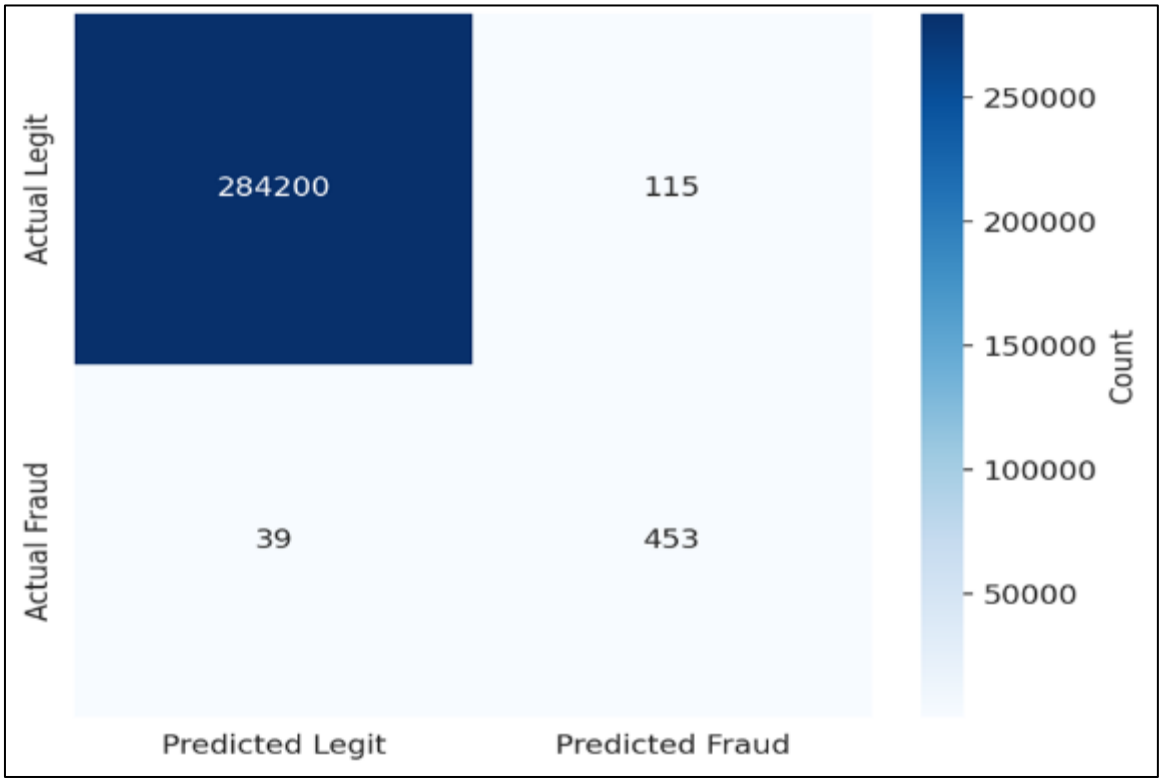


Figure 4: Confusion Matrix of the Hybrid CNN–XGBoost Model

4.3 Feature Importance and Training Dynamics

An advantage of retaining XGBoost as the final classifier is the availability of feature-importance scores. Several of the highest-ranked features are CNN-derived embeddings, confirming that the convolutional stage contributes discriminative information beyond the original tabular attributes. Amount, V14, V12, and Time also remain influential, consistent with domain knowledge of fraud indicators (Dal Pozzolo et al., 2015). Training and validation loss curves of the CNN feature extractor decrease smoothly and converge over 30 epochs, with only a modest gap between training and validation loss, indicating effective regularization and limited overfitting.

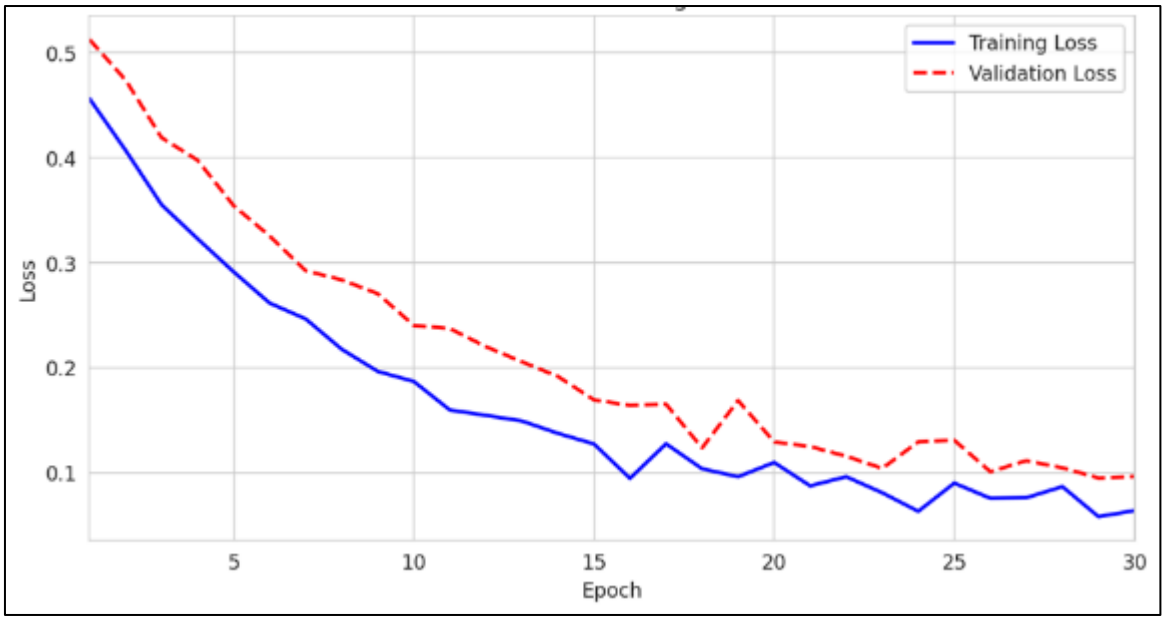


Figure 5: CNN Feature Extractor – Training and Validation Loss Curves

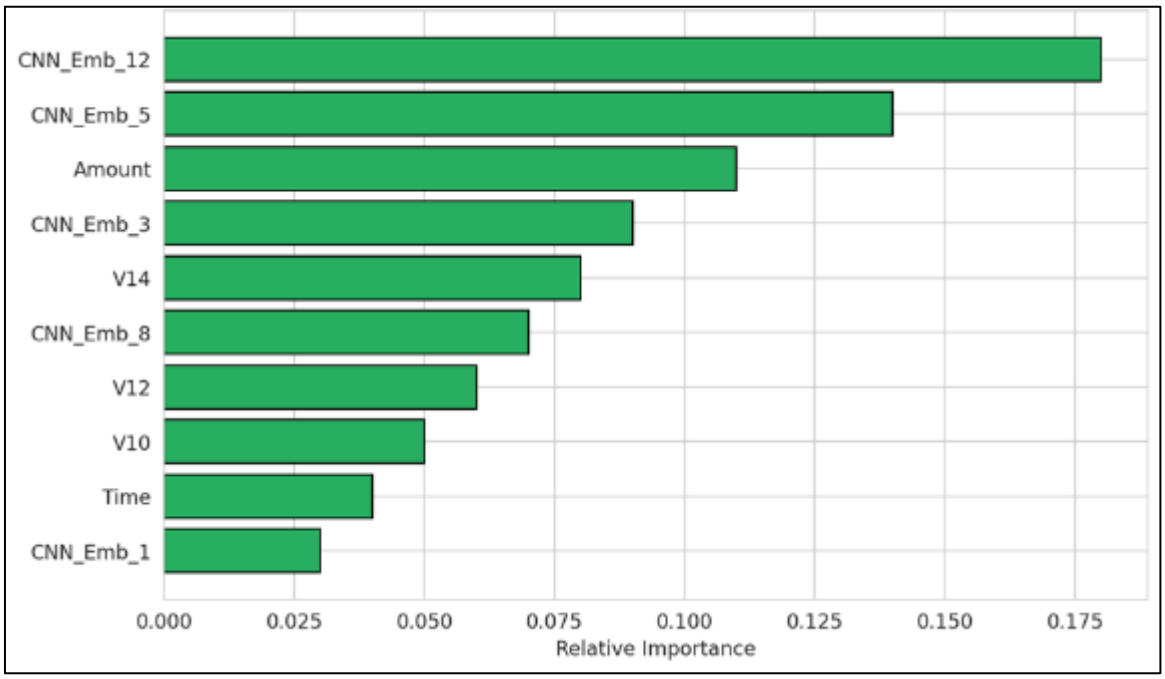


Figure 6: Top Feature Importance from the XGBoost Stage of the Hybrid Model

5 DISCUSSION

The experimental results demonstrate that the hybrid CNN–XGBoost architecture consistently outperforms both classical machine-learning baselines and the individual components of the hybrid itself. The performance gain relative to standalone XGBoost indicates that the CNN successfully extracts complementary hierarchical patterns that are not fully captured by hand-crafted or raw tabular features alone. Conversely, the improvement over the standalone CNN suggests that the regularised tree ensemble provides a more robust decision boundary under severe class imbalance than a purely neural classifier.

These findings align with recent literature reporting advantages of hybrid deep-learning and gradient-boosting pipelines for financial fraud detection (Chen et al., 2025). The low false-positive rate achieved by the hybrid model is particularly noteworthy from an operational perspective, as it implies reduced customer friction and lower manual-

review overhead while still maintaining high recall of fraudulent transactions. Limitations of the present evaluation include reliance on a single primary benchmark dataset and the use of offline batch inference rather than a live production stream. Future work should examine generalisation across additional digital channels (e.g., mobile-money platforms) and assess robustness under simulated concept drift.

6 CONCLUSION

This study has demonstrated that a hybrid architecture combining Convolutional Neural Networks for automatic hierarchical feature extraction with Extreme Gradient Boosting for final classification constitutes an effective approach to fraud detection in digital financial transactions. On the European Credit Card Fraud Detection benchmark the hybrid model achieved superior performance across accuracy (0.995), precision (0.930), recall (0.920), F1-score (0.925), AUC-ROC (0.988), and AUPRC (0.930), while recording the lowest false-positive rate (0.005). Feature-importance analysis confirmed that CNN-derived embeddings ranked among the most influential predictors, validating the contribution of the deep representation stage.

The findings reinforce the broader methodological principle that complementary model families—deep representation learners and regularised ensemble classifiers—can be fused to produce systems that outperform either component used in isolation. In the specific domain of digital payments, the demonstrated gains in recall, precision-recall metrics, and false-positive control have direct implications for loss reduction, operational efficiency, and customer experience.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

We sincerely thank the Department of Computer Science of the Falculty of Computing in Delta State University Abraka. We also appreciate the management of Federal Polytechnic Orogun, Delta State. To our family members, friends and well wishes, we say a big thank you for all the support and assistance rendered. Finally, to God Almighty we render all glory for making it possible for the above mentioned persons to give their impacts.

REFERENCES

[1] Alarfaj, F. K., Malik, I., Khan, H. U., Alomair, A., Almusallam, N., & Alshobaili, F. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 10, 39700–39715.

[2] Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review, 40, 100402.

[3] Baisholan, N., et al. (2025). Comparative analysis of machine learning models for credit card fraud detection under severe class imbalance. Journal of Financial Crime (forthcoming).

[4] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828.

[5] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613.

[6] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255.

[7] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[9] Chen, Y., et al. (2025). Deep learning approaches for financial fraud detection: A comprehensive survey (2019–2024). IEEE Transactions on Neural Networks and Learning Systems (forthcoming).

[10] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 159–166).

[11] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[13] Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In International Conference on Neural Information Processing (pp. 483–490). Springer.

[14] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

[15] Husejinović, A. (2020). Credit card fraud detection using improved deep learning models. International Journal of Computer Science and Network Security, 20(5), 1–8.

[16] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[17] Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In 28th European Modeling and Simulation Symposium (pp. 249–255).

[18] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569.

[19] Ravitch, S. M., & Riggan, M. (2017). Reason & rigor: How conceptual frameworks guide research (2nd ed.). SAGE Publications.

[20] Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence, 76, 130–157.

[21] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

[22] Sorinola, M., & Hamed, A. (2026). Digital financial transactions and the evolving landscape of payment fraud. Journal of Financial Technology (forthcoming).

[23] Vapnik, V. N. (1998). Statistical learning theory. Wiley.