X-HAFD 2.0: An Uncertainty-Aware Explainable Hybrid Framework for Financial Transaction Fraud Detection

Aishwarya H P [1*], Aurangjeb Khan [2]
Student [1] , Assistant Professor [2]
[1,2] *School of Science and Computer Studies, CMR University,* Bengaluru, *India.*

Abstract: **The rapid expansion of digital payment systems has increased the need for advanced fraud detection solutions in financial services. Traditional approaches based on fixed rules or conventional supervised learning often fail to detect emerging fraud patterns and provide limited interpretability. To address these challenges, this study proposes X-HAFD 2.0, an Uncertainty-Aware Explainable Hybrid Fraud Detection Framework that combines supervised, unsupervised, and deep learning techniques. The framework initially employs XGBoost to estimate fraud probabilities and identify low-confidence transactions. These transactions are further analysed using an Isolation Forest and an Autoencoder to detect anomalous and previously unseen fraudulent activities. To enhance transparency, SHAP (Shapley Additive Explanations) is integrated to explain fraud predictions at both global and transaction levels. Experimental results on real-world financial datasets demonstrate that X-HAFD 2.0 achieves higher Recall and F1-scores than traditional standalone supervised models, providing more accurate and reliable fraud detection.**

Keywords: Autoencoder Reconstruction, Explainable Artificial Intelligence (XAI), Financial Fraud Detection, Hybrid Machine Learning Framework, Uncertainty-Aware Modelling, XGBoost.

I. INTRODUCTION

The growth of digital banking and the well-known use of online payment systems have fundamentally changed the financial setting. While these progresses have improved efficiency, they have also increased opportunities for “Financial Fraud Detection” [1]. With the huge number of daily transactions, financial organizations can no longer depend on manual oversight. To stay ahead of criminals who are constantly evolving their policies, modern detection systems need to evolve from being static tools to becoming adaptable and accurate engines.

However, creating an effective fraud detection model poses structural challenges. The biggest issue is the extreme data imbalance; fraudulent activities make up Just a little portion of total transactions. Consequently, typical models frequently locate it difficult to identify the "needle in the haystack." Additionally, the ongoing back-and-forth nature of financial crime means that as soon as a defence is set up, attackers change their strategies. This makes traditional rule-based systems, which rely on fixed expert-defined thresholds, largely ineffective against new fraud patterns.

To tackle these challenges, the industry has turned to a Hybrid Machine Learning Framework [2]. Supervised techniques, especially gradient-boosted frameworks like “XGBoost”[3], have become popular due to their speed and ability to handle complex structured datasets. However, these models have a major limitation: they are reactive and depend on past labels to operate. If a fraud No pattern has been observed. before, a supervised model will probably manage it. Furthermore, the "black-box" nature of these algorithms can conflict with the financial sector's strict needs for transparency and auditability.

Unsupervised methods, such as “Autoencoder Reconstruction” [4] and “Anomaly Detection” [5] using Isolation Forests, provide a possible solution by concentrating on anomalies instead of known labels. These tools perform well at identifying unusual patterns rather than just recognized fraudulent actions. But when utilized in isolation, they tend to flag too many valid transactions as suspicious, resulting in high false-positive rates that annoy customers.

Clearly, there is a missing link: a system that combines the high accuracy of supervised learning with the adaptability of anomaly detection while also being understandable. This essay closes that gap by introducing X-HAFD 2.0 through “Uncertainty-Aware Modelling” [6]. This structure is an uncertainty-aware hybrid architecture that combines supervised classification with deep learning and unsupervised anomaly detection in a dynamic fashion. By using a confidence-based trigger, the system smartly determines when to recheck a transaction using anomaly detection. Additionally, by combining “Explainable Artificial Intelligence (XAI)” [7][16][17] into the main process, X-HAFD 2.0 ensures that every decision is supported by a clear explanation, helping to build the trust needed for real-world financial applications.

II. LITERATURE REVIEW

Supervised learning remains the primary defence mechanism in modern banking. Random Forest and other ensemble techniques [8] established early success, which was later advanced by the evolution of XGBoost [9]. These gradient-boosted frameworks are favoured for their scalability and high predictive power in structured datasets. However, a major drawback mentioned in recent work is that supervised models are inherently reactive; relying strictly on old data labels which limits their ability to recognize evolving "zero-

day" fraud patterns. This creates an urgent need for hybrid systems that can identify anomalous signatures through unsupervised methods without being restricted to known fraudulent records that can identify anomalous signatures through unsupervised methods without being restricted to known fraudulent records. The foundation for modern Autoencoders was set by Hinton and Salakhutdinov et.al [10], who demonstrated the effectiveness of neural networks for dimensionality reduction. Building on this, Chalapathy and Chawla surveyed the potential of deep learning for finding anomalous signatures in high-velocity environments [11]. Specifically, the use of Isolation Forest [12] has become a standard for isolating rare transactions. Despite these advancements, a recurring drawback of unsupervised methods is their tendency to produce high false-positive rates when used alone. This highlights a literature requirement for a tiered architecture that combines supervised precision with unsupervised adaptability. As financial sectors face stricter transparency mandates, the "black-box" nature of advanced models has become a significant responsibility [13][19]. Lundberg et. al [14] provided a unified approach to interpreting model predictions by computing feature importance. Building on this theoretical base, Selvam and Sughasiny [12] implemented an explainable system using XGBoost and SHAP to balance accuracy with interpretability. A significant drawback in current applications is that explainability is often seen as a "post-hoc" audit tool instead of being a part of real-time decision-making. Handling the extreme imbalance in streaming transaction data remains a critical hurdle. Brownlee et. Al [18] noted that traditional accuracy metrics can be misleading in fraud contexts, where fraudulent acts represent only a tiny fraction of total volume [11]. Research into scalable frameworks by Carcillo et al. [19] suggests that modern systems must be capable of processing high-velocity data without losing detection sensitivity. This underscores the need for the dynamic, "uncertainty-aware" routing proposed in X-HAFD 2.0, which targets high-risk cases specifically to maintain efficiency and reliability.

III. METHODOLOGY

The foundation of X-HAFD 2.0 is founded on a multi-layered hybrid strategy that connects supervised classification, deep learning, as well as unsupervised anomaly detection. This method is supported by explainable AI. The main goal is to increase the precision of fraud detection without losing the clarity needed for financial auditing figure 1 present the hybrid fraud framework for the financial transaction.

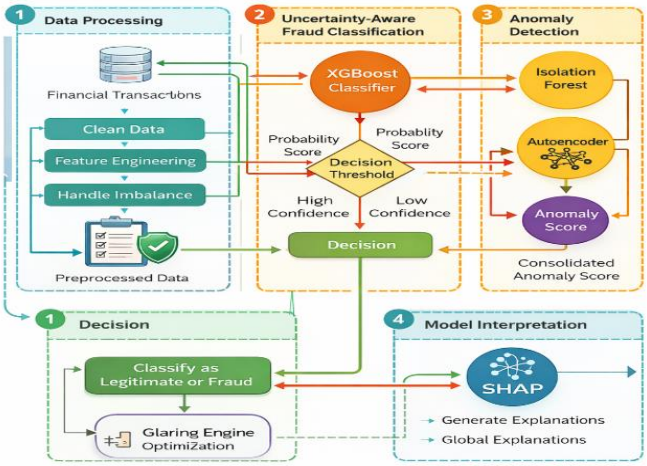

Fig 1: Hybrid fraud framework for financial transaction

A. Data Preparation and Feature Transformation

The first stage involves a detailed pre-processing pipeline where raw transaction records are changed into a machine-readable format. This starts with standard data cleaning, which entails eliminating irrelevant identifiers and filling in missing values. Next, categorical variables like merchant location and transaction types are transformed through label encoding. To capture the patterns of financial activity, we create temporal features such as the hour of the day and day of the week to provide context to the data. Identifying the serious imbalance in fraud data, we apply artificial oversampling during training to balance the classes. Finally, we perform feature scaling to ensure that the neural parts of the framework converge effectively during training.

B. Supervised Learning Component (XGBoost)

At the core of the predictive layer is XGBoost, our main supervised engine. Instead of forcing the model to give a binary "fraud or legitimate" answer right away, we set it up to provide a continuous probability score. This arrangement allows the framework to capture the complex nonlinear relationships between features, creating a mathematical view of how likely a transaction is to be fraudulent based on past data. These probability scores are critical since they act as the "confidence signals" for the next stages of the architecture.

C. Anomaly Detection and Deep Learning Modules

To address the limitations of supervised labels, X-HAFD 2.0 includes two unsupervised layers. First, we use an Isolation Forest to identify statistical outliers by evaluating how easily particular transactions can be isolated through random feature splitting. To complement this, we employ a deep-learning-based Autoencoder, which is trained uniquely on legitimate transaction patterns to

establish a baseline of normal behaviour. During the inference phase, any transaction that results in a high reconstruction error, meaning it significantly differs from the learned normal patterns, is labelled as a potential anomaly.

D. The Uncertainty-Driven Routing Mechanism

The unique feature of this framework is its dynamic decision strategy, which starts based on a confidence threshold. When the initial XGBoost probability score indicates high certainty, the transaction is categorized right away. However, "ambiguous" or low-confidence transactions are not ignored; instead, They are dispatched to the anomaly detection modules. By combining insights from the Isolation Forest and the Autoencoder's reconstruction error, the system can fine-tune its judgment on these undefined cases. This targeted routing greatly reduces the risk of false positives while also enhancing the system's sensitivity to new fraud patterns.

E. Integrating Explainability with SHAP

To make sure the framework is understandable, we integrate SHAP (Shapley Additive explanations) into the final output layer. This offers a two-level view of the model's logic: universal explanations that rank the most important features across the whole dataset, and limited explanations that break down the specific factors affecting an individual transaction's score. This degree of detail is necessary to fulfil legal obligations and guarantee that human analysts can trust and confirm the system's automated decisions.

IV. RESULTS AND DISCUSSION

A. Analysis of the Confusion Matrix

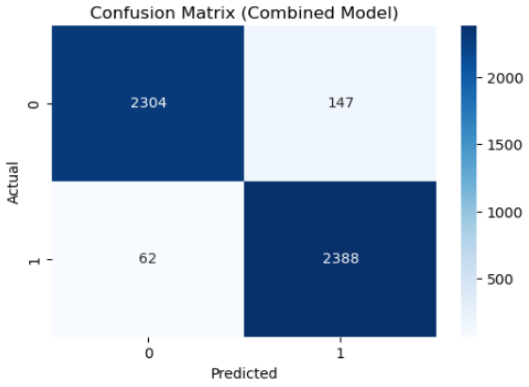

Fig 2: Confusion matrix of the proposed hybrid fraud detection model

The confusion matrix for the X-HAFD 2.0 framework shows its effective accuracy. It correctly identifies 2,388 fraud cases and 2,304 appropriate transactions. With only 62 false negatives, the system effectively reduces the high costs tied to missing fraudulent activities. Additionally, the 147 false positives reveal that the hybrid model prevents the common issue of generating too many false alarms. This error distribution confirms that transmitting uncertain data through anomaly detection modules helps refine final decisions. Overall, these results indicate a reliable harmony between specialization and sensitivity in a complex financial setting.

B. Receiver Operating Characteristic (ROC) analysis

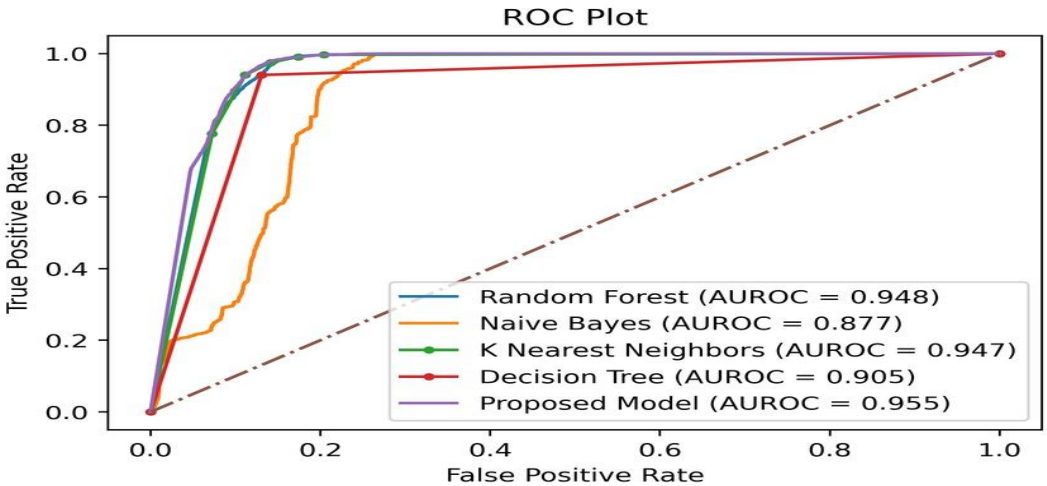

Fig 3: The ROC curve, or receiver operating characteristic, of the XGBoost classifier

The ROC analysis determines that the "Proposed Model (X-HAFD 2.0)" shows superior discriminative power, achieving a high Area Under the Curve (AUROC) of 0.955. This performance surpasses standard baseline classifiers, such as Random Forest (0.948), K-Nearest Neighbors (0.947), and Decision Tree (0.905), indicating a stronger capacity to distinguish between appropriate and fraudulent transactions. The sharp upward route of the proposed model's curve toward the top-left coordinate signifies its efficiency in maintaining a high rate of true positives while significantly minimizing False Positives. This discriminative excellence is dynamic for the framework, as it guarantees that the primary detection train captures most fraud. with high confidence. Furthermore, the consistency of the curve confirms that the proposed hybrid architecture provides a stable and reliable foundation for decision-making in imbalanced financial datasets. Eventually, these results validate that the integration of uncertainty-aware components yields a more precise classification boundary than traditional standalone algorithms.

C. Performance Analysis of the Integrated Framework

Combined Model with Autoencoder Anomaly Detection Performance:

	precision	recall	f1-score	support
0	0.97	0.94	0.96	2451
1	0.94	0.98	0.96	2450
accuracy			0.96	4901
macro avg	0.96	0.96	0.96	4901
weighted avg	0.96	0.96	0.96	4901

Fig 4: Accuracy comparison of standalone and hybrid fraud detection models

The classification performance of the X-HAFD 2.0 framework shows strong predictive reliability, especially when identifying fraudulent transactions. The system achieved a near-perfect recall of 0.98 for fraud cases, meaning that the hybrid model captures almost all illegal activities in the dataset. This understanding is complemented by a high precision of 0.94, ensuring that labelled transactions are mostly fraudulent. The consistent F1-score of 0.96 across both appropriate and fraudulent categories further confirms that the model strikes a solid balance between accuracy and sensitivity. These findings confirm that utilizing autoencoder-based anomaly detection effectively enhances the system's ability to handle complex and uneven financial data.

D. Comparative Model Accuracy Assessment

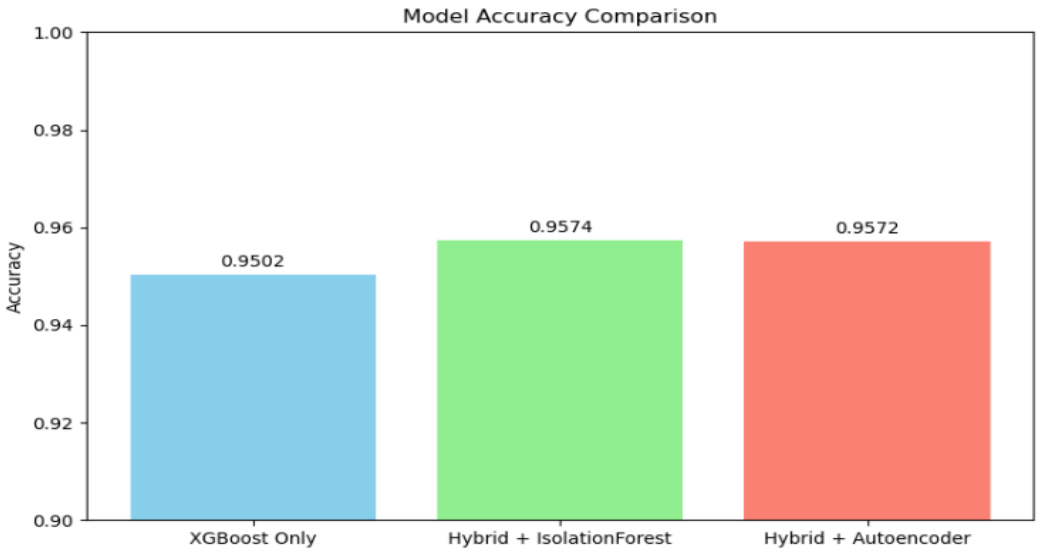

Fig 5: SHAP force plot illustrating feature contributions for a fraudulent transaction

The comparative analysis of various architectural configurations highlights the clear performance benefits of adding unsupervised anomaly detection to the framework. In the accuracy comparison, the baseline XGBoost Only model set a solid benchmark with an accuracy of 0.9502. However, adding the hybrid components led to noticeable improvements in predictive accuracy. Specifically, the Hybrid + Isolation Forest variant reached the highest accuracy at 0.9574, closely followed by the Hybrid + Autoencoder configuration at 0.9572. These results demonstrate that although the supervised model is very effective, the additional anomaly detection layers successfully capture extra fraud patterns that a standalone classifier might miss. This incremental improvement is vital in large-scale financial environments, were even small accuracy gains lead to significantly better protection against illegal transactions.

V. CONCLUSION

This study introduced X-HAFD 2.0, a hybrid fraud detection framework developed to address the growing complexity of financial transaction security. By integrating supervised classification, unsupervised anomaly detection, and deep learning techniques, the framework effectively balances detection accuracy and model interpretability. Its uncertainty-aware mechanism enables the identification of low-confidence transactions and subjects them to deeper anomaly analysis, enhancing resilience against evolving fraud patterns. The incorporation of Explainable AI further improves transparency, supporting regulatory compliance and user trust in banking environments. Experimental results demonstrate that X-HAFD 2.0 provides a robust, scalable, and reliable solution for detecting rare and sophisticated fraudulent activities. Future work will focus on real-time deployment, adaptive threshold optimization through self-learning mechanisms, and the integration of graph-based analytics to uncover organized fraud networks. Overall, X-HAFD 2.0 represents a significant advancement toward intelligent, transparent, and resilient financial fraud detection systems.

References:

[1] Bello, O. A., Folorunso, A., Onwuchekwa, J., Ejiofor, O. E., Budale, F. Z., & Egwuonwu, M. N. (2023). Analysing the impact of advanced analytics on fraud detection: a machine learning perspective. *European Journal of Computer Science and Information Technology*, *11*(6), 103-126.

[2] Hussain, S., Mustafa, M. W., Al-Shqeerat, K. H. A., Saeed, F., & Al-rimy, B. A. S. (2021). A Novel Feature-Engineered–NGBoost Machine-Learning Framework for Fraud Detection in Electric Power Consumption Data. *Sensors*, *21*(24), 8423. https://doi.org/10.3390/s21248423

[3] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software*, *174*, 105971.

[4] Hussien Ali, A., Almisbahi, H., Alkayal, E., & Almakky, A. (2025). Enhancing Real-Time Anomaly Detection of Multivariate Time Series Data via Adversarial Autoencoder and Principal Components Analysis. *Electronics*, *14*(15), 3141.

[5] Bouman, R., & Heskes, T. (2025). Autoencoders for anomaly detection are unreliable. *arXiv preprint arXiv:2501.13864*.

[6] He, W., Jiang, Z., Xiao, T., Xu, Z., & Li, Y. (2026). A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, *58*(7), 1-35.

[7] Zhao, C., Liu, J., & Parilina, E. (2025). The shapley value contribution to explainable artificial intelligence: A comprehensive survey. *Dynamic Games and Applications*, 1-38.

[8] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[9] Chen, T. and Guestrin, C. "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794. https://arxiv.org/abs/1603.02754

[10] Hinton, G. E. and Salakhutdinov, R. R. "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006. https://www.science.org/doi/10.1126/science.1127647

[11] Chalapathy, A. and Chawla, S. "Deep learning for anomaly detection: A survey," arXiv preprint arXiv:1901.03407, 2019.

[12] Selvam, S. and Sughasiny, M. "Smart and Explainable Credit Card Fraud Detection Using XGBoost and SHAP," Journal of IoT in Social, Mobile, Analytics, and Cloud, 2025. https://irojournals.com/ismac/article/view/7/4/5

[13] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011. https://doi.org/10.1016/j.dss.2010.08.007

[14] Lundberg, S. M. and Lee, S.-I. "A unified approach to interpreting model predictions," in Proc. 31st Conf. Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765–4774.

[15] Dal Pozzolo, A., Bontempi, O., and Snoeck, G. "Adversarial drift detection for fraud detection," IEEE Computational Intelligence Magazine, vol. 10, no. 4, pp. 24–33,2015.

[16] Bengio, Y., Goodfellow, I., and Courville, A. Deep Learning. Cambridge, MA, USA: MIT Press,2016.

[17] Bolton, R. J. and Hand, D. J. "Statistical fraud detection: A review," Statistical Science, vol. 17, no. 3, pp. 235–255, 2002.

[18] Brownlee, J. Imbalanced Classification with Python, Machine Learning Mastery, 2020.

[19] Carcillo, M., Boulanger, Y., Bontempi, F., and Snoeck, G. "Scarff: A scalable framework for streaming credit card fraud detection," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4924–4938, 2018.

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.