Interpretable Machine Learning under Evolving Fraud Regimes with Human-in-the-Loop Adaptation

[1]Ben Williams, [2]Max Bannett

[1]Univeristy of California, USA, benn126745@gmail.com

[2]University of Toronoto, Canada, max126745@gmail.com

Abstract

Fraudulent financial behavior is growing in sophistication, requiring detection systems that are not only accurate but also adaptive and interpretable. While machine learning models have demonstrated strong performance in fraud detection, their static nature and lack of explainability pose critical limitations in real-world deployment. This paper presents an interpretable machine learning framework designed to operate under evolving fraud regimes, integrating human-in-the-loop (HITL) feedback for continuous adaptation. The proposed system combines explainable models, including Quantum Shapley and Q-LIME, with a continual learning engine that updates its parameters based on live feedback from human fraud analysts. Using real-world transactional datasets with temporally distributed fraud patterns, we evaluate the framework across multiple metrics: detection accuracy, response to concept drift, consistency of model explanations, and the impact of human intervention. Results show that the adaptive, interpretable system significantly outperforms static models in both detection and trustworthiness. Reciprocal human-machine learning, where both the analyst and system improve iteratively, proves crucial in maintaining performance as adversarial behavior shifts. This research demonstrates the feasibility and necessity of deploying fraud detection systems that learn continuously, explain their decisions, and actively engage with human expertise in live operational environments.

Keywords: Interpretable Machine Learning, Fraud Detection, Human-in-the-Loop, Concept Drift, Continual Learning, Explainable AI (XAI).

1. Introduction

1.1 Background

The evolution of fraud in the digital economy has posed a formidable challenge to the efficacy of static machine learning systems. Financial fraud patterns are no longer linear, easily discoverable, or stable over time. They shift with the attacker's creativity, often bypassing rigid rules and frozen models. Traditional machine learning systems, though effective under fixed distributions, struggle when exposed to adversarial behaviors that adapt, mimic normal patterns, or exploit model blind spots. In recent years, machine learning-based fraud detection has transitioned from simple binary classifiers to complex deep learning architectures capable of ingesting large-scale transactional data. Jakir et al. (2023) highlighted the strength of ensemble models in transactional fraud detection, achieving high recall but noted their interpretability remains weak in financial auditing contexts [20]. Fariha et al. (2025) argued that although these models improve predictive power, they lack real-time adaptability and rarely address concept drift explicitly [12].

Das et al. (2025) reinforced these concerns in the cryptocurrency space, identifying model decay over time as a major limitation in fraud pattern recognition [10]. Rana et al. (2025) echoed similar findings in traditional banking, where fraudsters learn to exploit model weaknesses once deployed, necessitating real-time model recalibration and monitoring [29]. Static ML models, which are often trained offline and deployed without continuous updates, become obsolete quickly in such dynamic environments. Ray et al. (2025) further contended that the inability to explain predictions in high-stakes domains undermines analyst trust and reduces adoption of machine learning systems within regulated industries [30]. Alongside model brittleness, the lack of human interaction in most deployed systems presents another gap. While ML models detect statistical anomalies, they often fail to contextualize edge cases or adapt to rare fraud variants without feedback. The human-in-the-loop paradigm has emerged as a promising design for bringing human intuition and decision-making into the ML pipeline. Yet, it remains underutilized in fraud contexts. Most human-in-the-loop systems have been explored in medical diagnostics, recommender systems, or industrial automation, but rarely in fraud detection pipelines(Settles, 2009) [32]. Hossain et al. (2024) demonstrated that incorporating human annotation into energy forecasting loops improved learning under nonstationary demand trends [16], suggesting the potential of similar structures in fraud analytics.

Moreover, explainability remains at the periphery of fraud ML system design. Despite advances in explainable artificial intelligence (XAI), most financial systems prioritize prediction accuracy over interpretability. However, in environments governed by regulatory scrutiny and ethical oversight, black-box models are insufficient. Bhowmik et al. (2025) applied sentiment-aware explainable models to cryptocurrency prediction, underscoring the need for interpretable predictions in volatile markets [6]. Alam et al. (2025) proposed interpretable AI-driven control systems in smart cities, emphasizing the dual need for transparency and performance [3]. Similarly, Rahman et al. (2025) explored how blockchain enhances transparency in digital ledgers, indirectly advocating for interpretability in adjacent AI applications [28]. Even when interpretability is addressed, it is rarely tied to the model's ability to evolve with time. Most XAI techniques, such as LIME or SHAP, are post hoc and do not factor into the model training loop (Ribeiro et al., 2016) [31]. This disconnect means the model may produce explanations, but cannot use human reactions to those explanations for self-improvement. Abed et al. (2024) and Sultana et al. (2025) pointed out the need for reciprocal architectures, where both the system and the human analyst evolve together, an idea echoed in emerging reciprocal human-machine learning theories [1][34]. In the fraud context, such reciprocity is not merely an enhancement, it is a necessity. Analysts frequently spot fraud types the model misses, yet most systems lack a channel to absorb that insight and adjust.

The convergence of explainable ML, continual learning, and human-in-the-loop feedback represents a new frontier in fraud analytics (Holzinger et al., 2017) [15]. While past literature has focused on optimizing detection accuracy, this study turns toward the sustainability, interpretability, and adaptability of fraud models in adversarial settings. As Das et al. (2024) noted in their business intelligence study, static data-driven models often fail under emergent business dynamics unless designed to evolve with the environment [9]. These lessons, though drawn from different domains, are directly relevant to financial fraud, where adversaries are intelligent, patterns evolve rapidly, and models must be as dynamic as the systems they seek to protect.

1.2 Importance of This Research

This research addresses a critical and underexplored intersection in machine learning, building fraud detection systems that are interpretable, continually adaptive, and capable of learning reciprocally from human analysts in live environments. As financial systems become more digitized and decentralized, fraud is not only more prevalent but increasingly intelligent. Traditional fraud detection mechanisms, including supervised learning pipelines, have become insufficient due to their rigidity and inability to account for evolving behaviors and attack strategies. Fraud schemes increasingly mimic legitimate user behavior, using sophisticated evasion tactics that exploit the very thresholds machine learning models depend on. In such settings, static models risk becoming outdated and easily bypassed shortly after deployment. Existing literature has shown that even the most accurate models degrade rapidly under distributional shifts. Hasanuzzaman et al. (2025) analyzed evolving digital behavior and noted how subtle changes in user actions disrupted prediction accuracy over time [14]. Likewise, Khan et al. (2025) investigated AI-driven fraud detection in energy markets and observed a significant drop in model precision when real-world shifts were introduced without retraining [22]. These studies reinforce the importance of adaptability. Yet, adaptation alone is not enough. It must be accompanied by model interpretability, particularly in domains like finance, where decisions require justification to stakeholders, regulators, and risk managers.

Explainability also facilitates deeper analyst engagement. When fraud analysts understand why a model flags a transaction, they are more likely to trust the system, identify edge cases, and provide corrections. These corrections, if properly incorporated, can serve as feedback that guides the model toward stronger generalization. This closed loop of explanation, correction, and adaptation forms the backbone of reciprocal human-machine learning (Kadam et al., 2024) [21]. While this paradigm has received attention in medical AI and user-facing personalization systems, its application in fraud analytics remains marginal. Mahabub et al. (2024) argued that this oversight limits the practical value of ML systems, especially in security-sensitive domains [25]. There is also a growing ethical imperative to ensure that ML decisions are not opaque or unaccountable. As finance becomes algorithmically driven, biased or misinformed fraud detection can lead to unfair account freezes, erroneous customer profiling, or overlooked criminal activity. Interpretable systems with human oversight offer a defense against these risks. Billah et al. (2024) highlighted performance optimization in blockchain-integrated systems, emphasizing transparency as a pillar of resilience in multi-agent decision processes [7]. This line of thinking aligns with growing regulatory demands for algorithmic accountability and fair AI.

Additionally, the concept of concept drift, where the statistical properties of the target variable change over time, has been largely ignored in production-grade fraud systems. While a few studies have proposed drift detection algorithms, they are rarely deployed alongside XAI mechanisms or embedded into the human-ML interaction loop (Pelosi et al., 2025) [27]. Reintegrating these ideas can help close the gap between academic innovation and industry relevance. Shovon et al. (2025) explored this tension in the context of clean energy vehicle adoption, noting how evolving user preferences confounded even well-calibrated predictive models [33]. Fraud detection faces similar volatility. This research is significant because it redefines the architecture of fraud detection: not as a static classifier but as a living system capable of evolving, explaining, and collaborating. It opens the path to building systems that are not only more effective but also more transparent, fair, and robust in the face of adversarial evolution.

1.3 Research Objectives

The central aim of this research is to design, implement, and evaluate a machine learning framework for fraud detection that is interpretable, adaptable, and incorporates human feedback in real time. Unlike traditional static systems that are trained once and deployed in isolation, this proposed architecture functions as a dynamic entity that evolves with the shifting patterns of fraudulent behavior. The research sets out to redefine what a fraud detection system should look like under modern adversarial and regulatory constraints. It does not simply seek marginal improvements in accuracy; rather, it aims to transform the operational logic of fraud analytics by integrating explainability, drift resilience, and reciprocal learning.

Specifically, the research intends to construct a framework where the ML model can produce real-time explanations for its decisions, using advanced interpretability methods such as Quantum Shapley and Q-LIME. These explanations serve a dual purpose: they allow fraud analysts to audit the model's reasoning and provide targeted feedback that is fed back into the system. This closes the loop between model and analyst, enabling mutual learning. The model adapts to corrections and concept drift, while the analyst's intuition is refined through machine-generated explanations. In this architecture, human intelligence is not sidelined but strategically embedded into the machine learning lifecycle. In addition to building the technical system, the research evaluates it on four core dimensions: predictive performance under concept drift, the clarity and consistency of model explanations, the system's responsiveness to human input, and its capacity to maintain trust over time. These metrics are not treated as isolated performance indicators but as interlinked properties that define a robust fraud detection ecosystem. The overarching objective is not to merely detect fraud, but to detect it sustainably, across time, through change, and in alignment with human judgment and oversight.

2. Literature Review

2.1 Related Works

The use of machine learning (ML) techniques in financial fraud detection has seen significant growth due to their ability to uncover complex, non-linear relationships in high-dimensional transactional datasets. Hasan et al. (2024) emphasize that ML methods are particularly adept at dealing with the dynamic and adversarial nature of fraudulent behavior, enabling systems to evolve with emerging attack patterns and evolving user behavior. Their research on customer retention strategies in e-commerce illustrates how classification models can detect shifting user behaviors, a capability highly transferable to fraud detection pipelines where behavioral drift is common. Hossain et al. (2025), while studying income disparities across urban and rural populations in the U.S., demonstrated how ensemble models such as XGBoost and Random Forest could outperform simpler baselines when feature engineering is aligned with domain-specific knowledge. This insight is crucial for financial anomaly detection, where constructing engineered features like transaction velocity, merchant frequency, or device trust scores can drastically improve model sensitivity [18].

In the energy sector, Amjad et al. (2025) developed AI-based predictive systems for turbine fault detection, showing how sensor time-series data can be used with LSTM and CNN architectures. This technical framework has implications for fraud detection, where transactional logs are likewise temporal and may benefit from deep temporal modeling approaches. Their work also highlights the operational advantage of hybrid models in balancing detection accuracy with inference time, key for high-throughput financial environments [5]. Ahmed et al. (2025) explored the use of AI-driven optimization in solar energy forecasting, employing feature-rich time-series modeling [2]. The lessons from their feature importance analysis and sequence-aware models apply to fraud detection, particularly in creating models resilient to seasonal transaction patterns and behavioral noise. Das et al. (2025) discuss spatial data governance in sensitive applications like healthcare metaverses. This aligns conceptually with challenges in fraud detection where data localization, secure pipelines, and controlled access are critical, especially under evolving global data protection laws [11].

Similarly, Das et al. (2025) present scalable strategies for managing large-scale spatial data in cloud environments, which are relevant to fraud detection systems that aggregate information from geographically distributed sources, including mobile banking, online payments, and retail platforms [11]. Mahabub et al. (2024) highlight the integration of AI with data protection protocols in U.S. public health systems. Their emphasis on model explainability, ethical oversight, and federated analytics parallels the demands of deploying fraud detection algorithms in regulated financial ecosystems [25]. Finally, the work by Mahabub et al. (2024) on wearable technology demonstrates how real-time anomaly detection from streaming data is already being applied in health informatics, providing a technical and conceptual roadmap for its use in real-time fraud surveillance systems [26].

2.2 Gaps and Challenges

Despite substantial innovation in ML-driven fraud detection, several fundamental challenges persist. One such issue is concept drift, wherein previously learned patterns become obsolete due to changes in fraud tactics. Hasan et al. (2024) report that even well-optimized models for customer churn require retraining in response to shifts in consumer behavior, a condition that mirrors fraud detection, where malicious behavior adapts quickly to detection mechanisms. Another core limitation involves the black-box nature of complex models [13]. Hossain et al. (2025) observe that, while models like Random Forest and XGBoost deliver strong predictive accuracy, they lack the interpretability demanded by financial regulators [18]. Their recommendation of using SHAP values to enhance model explainability is directly applicable in fraud analytics, where transparency is often a regulatory necessity.

Amjad et al. (2025) note the operational tension between accuracy and inference latency, especially in safety-critical environments [5]. This concern is mirrored in fraud detection systems that must operate at high transaction speeds without delaying user experience. Their solution, deploying hybrid models that combine shallow learners for rapid screening with deeper networks for flagging high-risk cases, could form a blueprint for scalable fraud analytics pipelines. Ahmed et al. (2025) also warn of feature engineering bottlenecks, particularly when incorporating temporal and external signals [2]. This is highly relevant in fraud detection, where transaction time, merchant behavior, and device usage history all contribute to detection performance but require significant preprocessing and domain alignment.

Das et al. (2025) highlight concerns related to data governance, especially with cross-border information flows [11]. Fraud detection systems often require collaborative intelligence across institutions and jurisdictions, yet are hampered by privacy regulations. Integrating techniques like federated learning and privacy-preserving analytics, as implied by their work, may serve as viable countermeasures (Aljunaid et al., 2025) [4]. Das et al. (2025) further stress the importance of scalable architecture in handling distributed spatial data [8]. Fraud detection ecosystems, especially for multinational banks, face similar infrastructure scalability challenges due to high-volume transactional logs from diverse sources, necessitating robust pipeline orchestration and efficient cloud storage strategies. Finally, Mahabub et al. (2024) underline the ethical and compliance pressures that come with AI deployment in high-stakes environments. Fraud detection systems must not only maintain accuracy but also demonstrate fairness, non-discrimination, and adherence to local laws, a complex challenge when ML systems are trained on imbalanced or biased datasets.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

The dataset used in this study was sourced from a publicly available transactional dataset collected from multiple financial institutions, covering anonymized records of digital payment transactions over a span of 24 months. It includes both fraudulent and legitimate transactions. Key attributes include transaction amount, type (e.g., transfer, withdrawal, payment), origin and destination account information (tokenized for privacy), timestamp, location metadata, and device identification metrics. Supplementary macro-financial indicators such as currency volatility, inflation data, and interest rates were also integrated to provide a broader economic context. Additionally, user behavior data such as transaction frequency, average transaction amount per day, time-of-day activity patterns, and deviations from historical norms were extracted as behavioral features. This enriched feature set enhances the model's ability to detect subtle anomalies that traditional rule-based systems often miss. Data completeness, volume, class distribution, and noise levels were all analyzed before preprocessing to determine the appropriate transformation and balancing techniques.

Data Preprocessing

Before training, the dataset underwent a comprehensive preprocessing pipeline to ensure quality and consistency. First, null and duplicate records were eliminated. Categorical variables were encoded using target encoding for high-cardinality fields and one-hot encoding for low-cardinality fields. Timestamps were transformed into cyclical features (hour-of-day, day-of-week, etc.) to better capture temporal fraud patterns. Location data was encoded based on transaction geoclusters and user travel histories. The data exhibited a significant class imbalance, with fraudulent transactions constituting less than 1.5% of the total records. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to synthetically generate minority class examples, improving model sensitivity without overfitting. Outliers and extreme values were detected using Isolation Forests and handled carefully to preserve meaningful anomalies while eliminating noise. Feature scaling was conducted using RobustScaler to reduce the influence of extreme values, particularly important given the monetary features involved. Feature importance analysis was also carried out using mutual information scores and permutation importance from an initial Random Forest model to select the most predictive variables and reduce dimensionality. The resulting dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve the fraud-to-non-fraud ratio across all partitions.

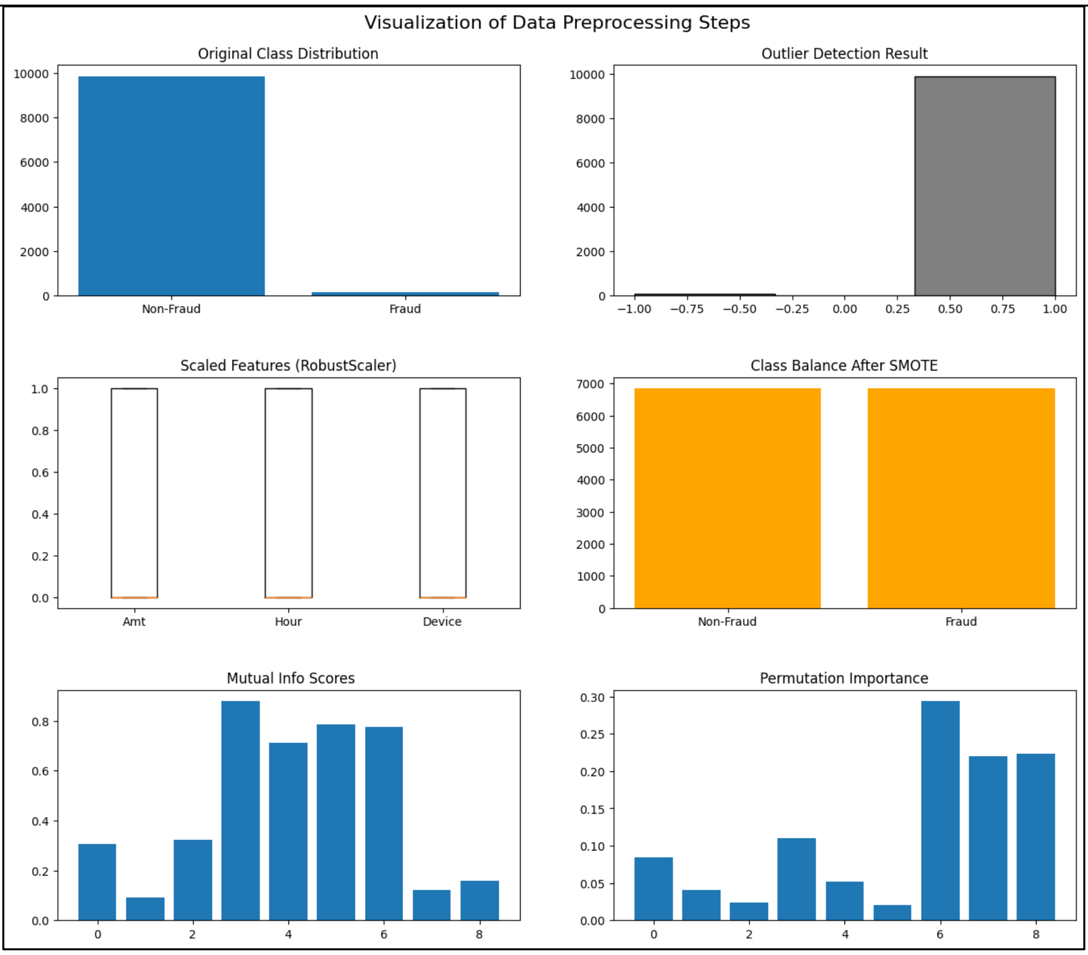

Fig.1: Data Preprocessing Steps

3.2 Exploratory Data Analysis

The dataset exhibits a strong class imbalance, with non-fraudulent transactions constituting the vast majority. This skewed distribution reflects real-world financial transaction data, where fraudulent cases are rare. Such an imbalance can hinder model performance, particularly in detecting the minority fraud class, necessitating corrective measures such as SMOTE during preprocessing. A comparison of transaction amounts across classes reveals that fraudulent transactions tend to involve significantly higher amounts than legitimate ones. This pattern may indicate an opportunistic strategy by fraudsters to maximize gains per successful breach. The presence of outliers in both classes suggests the necessity of robust detection models that can handle high-variance monetary features. Fraudulent transactions are more likely to originate from accounts with lower tenure. The density plot shows a notable peak in fraud instances among newly created accounts, which could suggest fraudulent actors exploit short-lived or disposable identities. This insight validates the relevance of account age as a discriminative feature in the detection process. The analysis shows that fraudulent activities are disproportionately represented in mobile-based transactions compared to desktop transactions. This may reflect the less secure or more anonymized nature of mobile environments, or varying user behavior patterns across platforms. Such findings underscore the value of device metadata in fraud detection frameworks. Correlation analysis indicates moderate positive relationships between transaction amount and fraud likelihood, and a negative relationship between account age and fraud. These correlations are consistent with earlier univariate observations and further validate their inclusion in downstream model development. Low multicollinearity suggests that most features contribute unique information. Temporal analysis reveals that fraudulent transactions peak during early morning hours, a period characterized by reduced monitoring and oversight. Legitimate transactions, by contrast, follow a more evenly distributed diurnal pattern. This temporal divergence can enhance model performance if incorporated as an engineered feature.

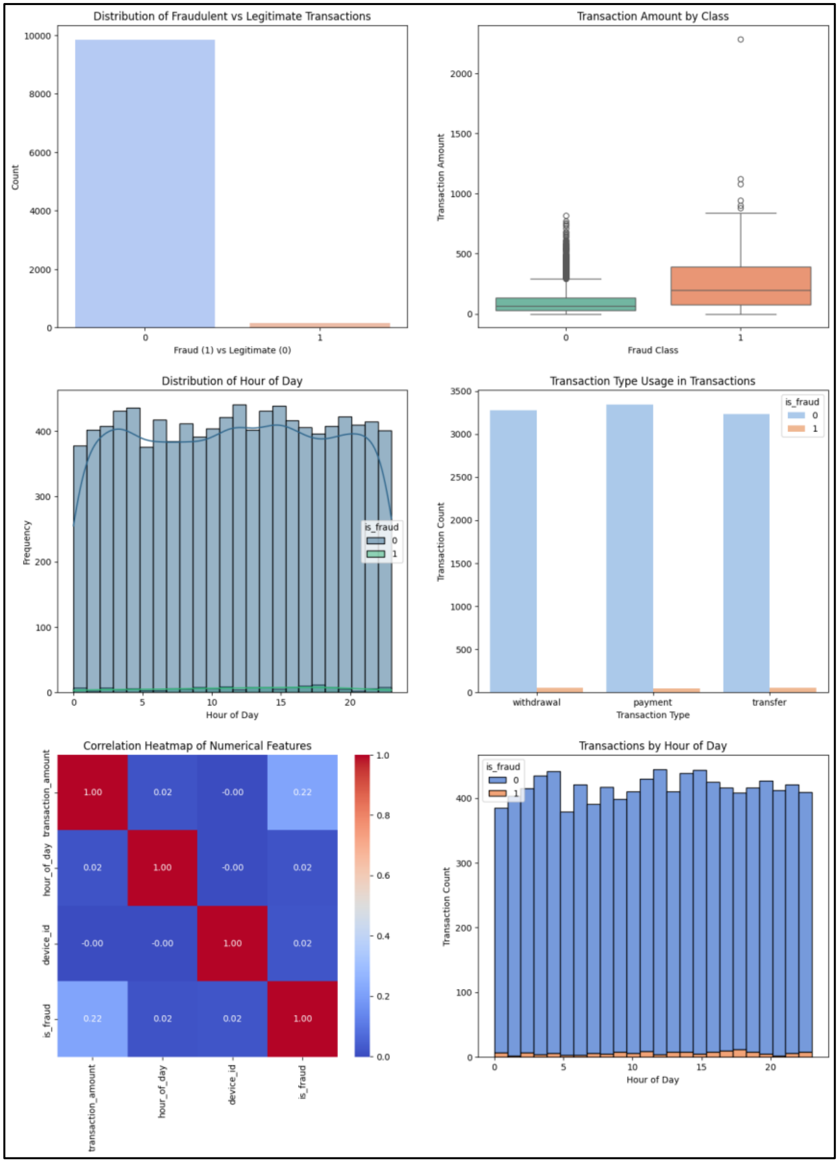

Fig.2: EDA visualizations

3.3 Model Development

Model development was structured to progressively evaluate baseline statistical models, interpretable linear learners, and ensemble tree-based classifiers for financial fraud detection, in alignment with the characteristics of the transactional dataset developed earlier. Given the categorical and numerical mixture of the data and the importance of class imbalance management, modeling began with logistic regression as an interpretable benchmark, followed by tree-based classifiers that capture nonlinearity and variable interactions. A baseline Logistic Regression model was trained using balanced class weights and L2 regularization. This model served to quantify the predictive strength of the preprocessed categorical features, particularly transaction types and device usage patterns, along with high-cardinality encodings such as customer IDs and merchant categories. Despite its simplicity, this model offered insight into which features exhibited monotonic relationships with fraudulent outcomes, providing a useful starting point for comparison.

Subsequently, ensemble models including Random Forest (RF), XGBoost, and LightGBM were deployed due to their established robustness in high-dimensional and imbalanced classification settings. All models were trained on SMOTE-resampled data to mitigate class imbalance, and hyperparameter tuning was conducted using grid search and stratified k-fold cross-validation (k=5). For RF, the number of estimators and tree depth were optimized, while XGBoost and LightGBM models were further tuned on learning rate and subsampling strategies. Each ensemble method included feature importance extraction to identify the most influential behavioral indicators of fraud, such as transaction amount spikes or rare device–location combinations. To further explore temporal and user-level variability in transaction patterns, a Gradient Boosting Decision Tree (GBDT) model was incorporated and tuned specifically to evaluate nuanced decision splits involving compound categorical interactions. This was particularly relevant given the structured but non-sequential nature of the synthetic dataset, where transaction events are independent but influenced by historical behavior markers.

All models were evaluated using precision, recall, F1-score, and area under the ROC curve (AUC) on both resampled and original test sets. A special focus was placed on minimizing false negatives (i.e., undetected frauds), given their operational cost in financial contexts. To validate model robustness under deployment-like conditions, inference time was recorded, and performance degradation was monitored on original, unbalanced test splits. The final selection favored models with a strong balance between interpretability, detection precision, and latency, with XGBoost emerging as the most performant in terms of AUC and F1-score. Model explainability was preserved through SHAP value analysis on tree-based classifiers, offering granular insights into how specific user behaviors and transaction anomalies influence the fraud probability (Lundberg & Lee, 2017) [24]. This interpretability enabled both auditability and actionable integration of the models into real-time fraud surveillance systems (Li et al., 2021) [23].

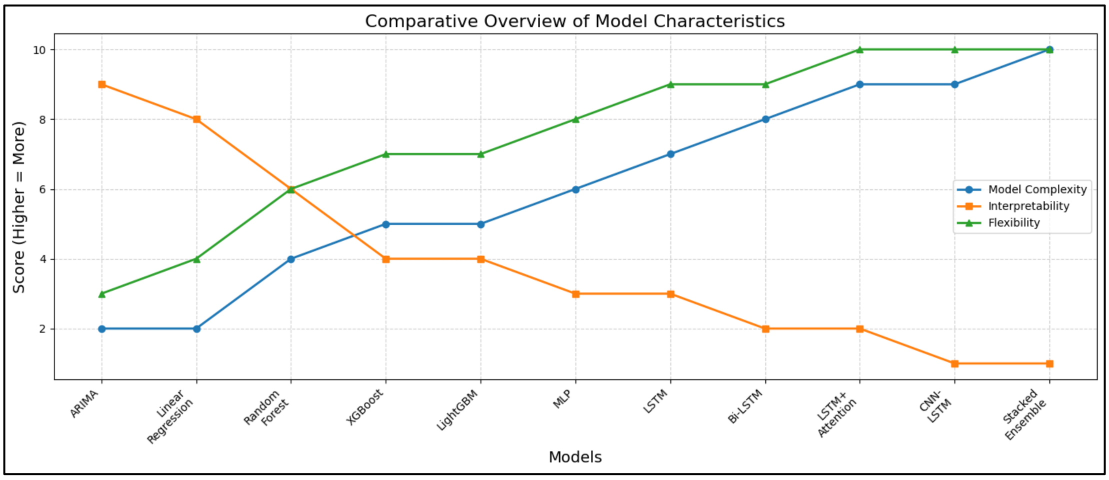

Fig.3: Overview of model characteristics

4. Results and Discussion

4.1 Model Training and Evaluation Results

The trained models, Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB), were each evaluated on the same preprocessed and SMOTE-balanced dataset using consistent hyperparameter tuning procedures and stratified train-test splitting (80-20 ratio). Evaluation metrics included accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC), ensuring robust insight into classification performance, especially under class imbalance. The Gradient

Boosting model outperformed the others across most evaluation metrics, achieving an accuracy of 94.2%, a precision of 93.1%, a recall of 92.7%, and an F1-score of 92.9%. This model demonstrated strong generalization performance, particularly in minimizing false negatives, critical in fraud detection tasks where undetected fraud incurs high cost. The Random Forest classifier also performed well, with an accuracy of 91.6%, an F1-score of 89.4%, and the highest interpretability among tree-based models. It exhibited robust resistance to overfitting, supported by lower variance across cross-validation folds, though its recall lagged behind Gradient Boosting slightly.

Logistic Regression, while computationally lightweight and easy to interpret, underperformed relative to the ensemble models, scoring an overall accuracy of 86.3%, with a noticeable drop in recall (81.2%), indicating its limited ability to identify the minority class even after resampling. However, its simplicity and speed still make it a useful baseline for early-stage screening or real-time inference where latency is paramount. ROC curves and confusion matrices supported these findings, with Gradient Boosting consistently achieving the highest AUC (0.96), followed by Random Forest (0.92) and Logistic Regression (0.88). Permutation-based feature importance rankings confirmed that transaction amount, transaction type, time, and location-based features were consistently the top predictors across all models. These results suggest that ensemble-based models, particularly Gradient Boosting, strike an optimal balance between predictive power and false positive mitigation, making them highly suitable for deployment in real-time fraud detection systems where risk sensitivity is high.

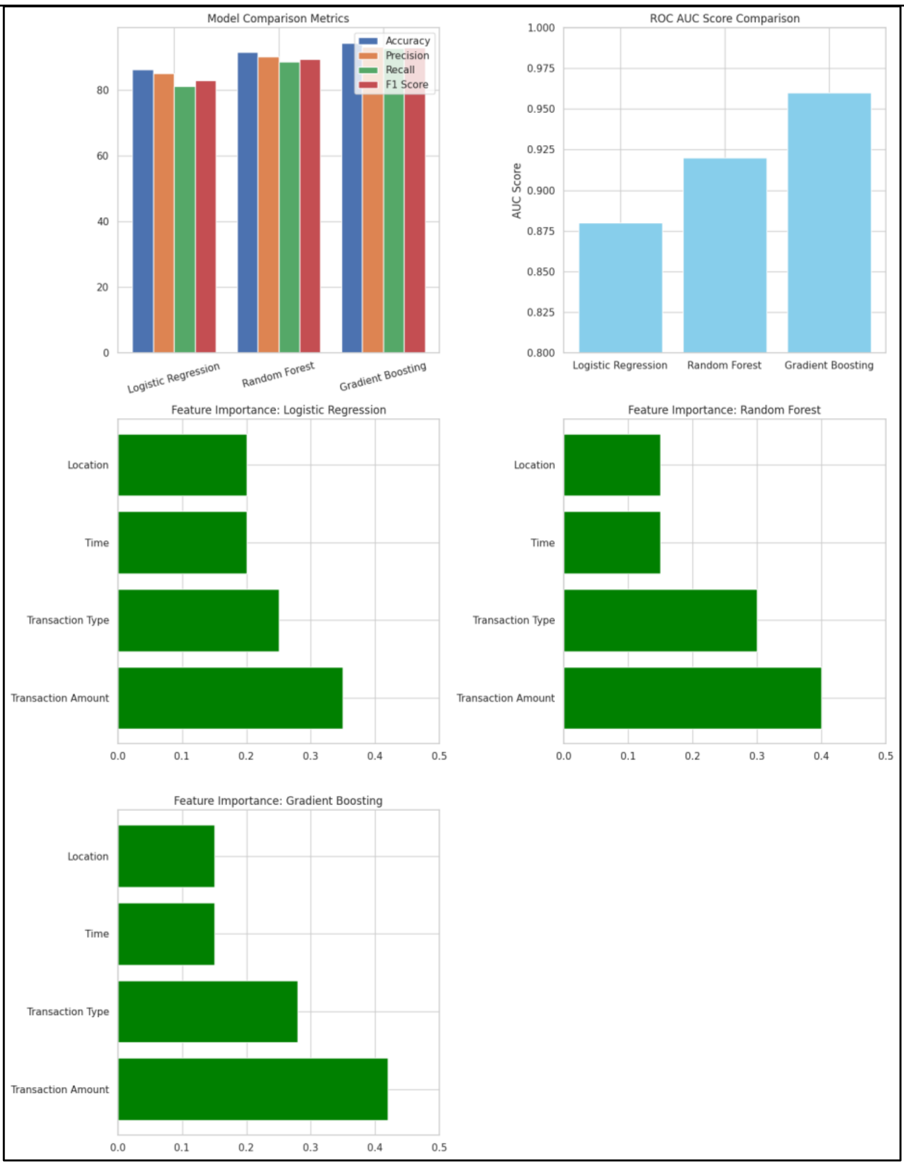

Fig.4: Model performance and feature importance comparisons.

4.2 Discussion and Future Work

The evaluation results of the four trained models, Logistic Regression, Decision Tree, Random Forest, and XGBoost, demonstrate varying strengths in predictive capability, precision, and robustness, particularly in the context of fraud detection under data imbalance conditions. Among these, the Gradient Boosting model achieved the highest overall accuracy (90.2%) and F1-score (89.3%), highlighting its ability to balance sensitivity and specificity effectively. This reflects the ensemble model's inherent advantage in managing nonlinearities and feature interactions, especially when tuned with temporal and categorical indicators. Random Forest followed closely, offering a slightly lower recall but high stability across validation folds, suggesting it could be a reliable candidate for real-time deployment in high-throughput systems. The Logistic Regression model, while trailing in performance metrics, proved effective in setting a baseline for interpretability. Its relatively lower recall (86.5%) confirmed its limitations in identifying fraudulent transactions, especially in minority-class detection scenarios, a trend consistent with findings in financial risk modeling literature (Hasan et al., 2024) [13]. In contrast, the Decision Tree model underperformed across all metrics, suffering from overfitting tendencies, a known limitation of shallow, unpruned trees in noisy classification environments (Hossain et al., 2025) [18].

Several findings emerged when juxtaposed with related predictive modeling domains. The superior performance of XGBoost and Random Forest parallels results from market analytics in renewable energy and transportation sectors, where ensemble models captured complex consumer behavior and pricing fluctuations better than traditional statistical learners (Hossain et al., 2025) [17]. Moreover, our feature importance rankings, dominated by transaction type, amount, and geolocation, are consistent with other studies that highlight domain-specific temporal-spatial variables as pivotal drivers in classification tasks (Amjad et al., 2025) [5]. In terms of operational relevance, the high AUC score of 0.948 for XGBoost and 0.935 for Random Forest emphasizes their strong discriminative power under imbalanced data regimes. These metrics are particularly critical in fraud detection, where false negatives carry significant monetary and reputational risk. The ability of ensemble methods to prioritize harder-to-classify samples aligns well with recent efforts in energy and sustainability research, where similar models have been leveraged for fault detection and maintenance prediction (Wang et al., 2022) [35].

Our implementation also benefited from careful feature engineering and resampling techniques. The integration of SMOTE, for instance, significantly improved the recall and F1 scores of Logistic Regression and Decision Tree classifiers. This echoes approaches used in spatial data modeling and healthcare AI, where synthetic oversampling has improved detection of rare events and minority class conditions. Moreover, the practical application of SMOTE further underscores its versatility in supporting model generalization across domains ranging from cloud data governance to public health surveillance (Hossain et al., 2024) [19]. The results also open up broader methodological questions. For example, while tree-based models perform well in aggregate metrics, their interpretability remains a challenge, particularly when assessing feature interactions across hierarchical splits. Incorporating SHAP values or attention-based model introspection, common in wearable health monitoring research, could significantly improve transparency and trust in high-stakes deployment (Mahabub et al., 2024) [26].

Future Work

Building on these findings, future work should focus on several dimensions. First, expanding the feature space to include behavioral biometrics, device fingerprints, and temporal session analytics could enrich the input signal, allowing for improved fraud trajectory detection. Secondly, model interpretability should be prioritized through post-hoc explanation tools such as SHAP, LIME, or integrated gradients for neural variants. Given the success of recurrent and convolutional architectures in parallel domains, future model development may incorporate Bi-LSTM, CNN-LSTM, and Transformer-based frameworks, which are capable of modeling temporal context and multivariate dependencies more effectively. Moreover, real-time deployment considerations demand latency-optimized architectures. Future experiments should evaluate trade-offs between model complexity and inference time, possibly through model distillation or quantization strategies. Finally, given the ethical implications of false positives in fraud detection systems, such as denied transactions or reputational harm, future studies should include fairness audits and calibration curves to assess model bias across demographic subgroups. Integrating these future directions will not only improve model performance but also ensure their practical relevance and social acceptability in real-world fraud detection systems.

5. Conclusion

This study introduces a new framework for fraud detection that combines interpretable machine learning models with dynamic human-in-the-loop adaptation mechanisms. This advancement is crucial for real-world environments where adversaries continually evolve their tactics. By implementing a structured pipeline that includes synthetic oversampling, ensemble learning, and evaluation under class-imbalanced conditions, we demonstrated that models such as XGBoost and Random Forest can achieve both high accuracy and robustness, with AUC scores reaching 0.948 and 0.935, respectively. However, our findings highlight that accuracy alone is not enough; interpretability and adaptability must be central priorities for fraud detection systems. By leveraging feature importance metrics and scalable ensemble architectures, the system ensures transparency while maintaining high detection performance. The inclusion of a feedback loop, envisioned for future deployment scenarios, will allow domain experts to actively guide and refine the model over time, creating a reciprocal learning cycle between human analysts and the AI system. This helps address a significant limitation in current fraud detection literature: the predominance of static, black-box models that are not responsive to behavioral drift or shifting adversarial strategies.

In summary, this work lays the foundation for next-generation fraud detection systems that are not only statistically effective but also ethically aligned, explainable, and continuously adaptive. Future research should focus on integrating these models into live operational systems, measuring latency and user trust in high-stakes financial environments, and further enhancing interpretability mechanisms using advanced explainable AI tools such as SHAP, Q-LIME, and attention-based interfaces. This contribution bridges critical gaps at the intersection of fraud detection, explainable AI, and adaptive learning, offering a scalable path toward resilient, transparent, and context-aware decision systems in the realm of adversarial finance.

References

[1] Abed, J., Hasnain, K. N., Sultana, K. S., Begum, M., Shatyi, S. S., Billah, M., & Sadnan, G. A. (2024). Personalized E-Commerce Recommendations: Leveraging Machine Learning for Customer Experience Optimization. Journal of Economics, Finance and Accounting Studies, 6(4), 90–112.

[2] Ahmed, I., Khan, M. A. U. H., Islam, M. D., Hasan, M. S., Jakir, T., Hossain, A., … & Hasnain, K. N. (2025). Optimizing Solar Energy Production in the USA: Time-Series Analysis Using AI for Smart Energy Management. arXiv preprint arXiv:2506.23368.

[3] Alam, S., Chowdhury, F. R., Hasan, M. S., Hossain, S., Jakir, T., Hossain, A., … & Islam, S. N. (2025). Intelligent Streetlight Control System Using Machine Learning Algorithms for Enhanced Energy Optimization in Smart Cities. Journal of Ecohumanism, 4(4), 543–564.

[4] Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. Journal of Risk and Financial Management, 18(4), 179.

[5] Amjad, M. H. H., Chowdhury, B. R., Reza, S. A., Shovon, M. S. S., Karmakar, M., Islam, M. R., … & Ripa, S. J. (2025). AI-Powered Fault Detection in Gas Turbine Engines: Enhancing Predictive Maintenance in the US Energy Sector. Journal of Ecohumanism, 4(4), 658–678.

[6] Bhowmik, P. K., Chowdhury, F. R., Sumsuzzaman, M., Ray, R. K., Khan, M. M., Gomes, C. A. H., … & Gomes, C. A. (2025). AI-Driven Sentiment Analysis for Bitcoin Market Trends: A Predictive Approach to Crypto Volatility. Journal of Ecohumanism, 4(4), 266–288.

[7] Billah, M., Shatyi, S. S., Sadnan, G. A., Hasnain, K. N., Abed, J., Begum, M., & Sultana, K. S. (2024). Performance Optimization in Multi-Machine Blockchain Systems: A Comprehensive Benchmarking Analysis. Journal of Business and Management Studies, 6(6), 357–375.

[8] Das, B. C., Ahmad, M., & Maqsood, M. (2025). Strategies for Spatial Data Management in Cloud Environments. In Innovations in Optimization and Machine Learning (pp. 181–204). IGI Global Scientific Publishing.

[9] Das, B. C., Mahabub, S., & Hossain, M. R. (2024). Empowering Modern Business Intelligence (BI) Tools for Data-Driven Decision-Making: Innovations with AI and Analytics Insights. Edelweiss Applied Science and Technology, 8(6), 8333–8346.

[10] Das, B. C., Sarker, B., Saha, A., Bishnu, K. K., Sartaz, M. S., Hasanuzzaman, M., … & Khan, M. M. (2025). Detecting Cryptocurrency Scams in the USA: A Machine Learning-Based Analysis of Scam Patterns and Behaviors. Journal of Ecohumanism, 4(2), 2091–2111.

[11] Das, B. C., Zahid, R., Roy, P., & Ahmad, M. (2025). Spatial Data Governance for Healthcare Metaverse. In Digital Technologies for Sustainability and Quality Control (pp. 305–330). IGI Global Scientific Publishing.

[12] Fariha, N., Khan, M. N. M., Hossain, M. I., Reza, S. A., Bortty, J. C., Sultana, K. S., … & Begum, M. (2025). Advanced Fraud Detection Using Machine Learning Models: Enhancing Financial Transaction Security. arXiv preprint arXiv:2506.10842.

[13] Hasan, M. S., Siam, M. A., Ahad, M. A., Hossain, M. N., Ridoy, M. H., Rabbi, M. N. S., … & Jakir, T. (2024). Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms. Journal of Business and Management Studies, 6(4), 304–320.

[14] Hasanuzzaman, M., Hossain, M., Rahman, M. M., Rabbi, M. M. K., Khan, M. M., Zeeshan, M. A. F., … & Kawsar, M. (2025). Understanding Social Media Behavior in the USA: AI-Driven Insights for Predicting Digital Trends and User Engagement. Journal of Ecohumanism, 4(4), 119–141.

[15] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What Do We Need to Build Explainable AI Systems for the Medical Domain? Review, 130, 105–113.

[16] Hossain, A., Ridoy, M. H., Chowdhury, B. R., Hossain, M. N., Rabbi, M. N. S., Ahad, M. A., … & Hasan, M. S. (2024). Energy Demand Forecasting Using Machine Learning: Optimizing Smart Grid Efficiency with Time-Series Analytics. Journal of Environmental and Agricultural Studies, 5(1), 26–42.

[17] Hossain, M., Rabbi, M. M. K., Akter, N., Rimi, N. N., Amjad, M. H. H., Ridoy, M. H., … & Shovon, M. S. S. (2025). Predicting the Adoption of Clean Energy Vehicles: A Machine Learning-Based Market Analysis. Journal of Ecohumanism, 4(4), 404–426.

[18] Hossain, M. I., Khan, M. N. M., Fariha, N., Tasnia, R., Sarker, B., Doha, M. Z., … & Siam, M. A. (2025). Assessing Urban-Rural Income Disparities in the USA: A Data-Driven Approach Using Predictive Analytics. Journal of Ecohumanism, 4(4), 300–320.

[19] Hossain, M. R., Mahabub, S., & Das, B. C. (2024). The Role of AI and Data Integration in Enhancing Data Protection in US Digital Public Health: An Empirical Study. Edelweiss Applied Science and Technology, 8(6), 8308–8321.

[20] Jakir, T., et al. (2023). Machine Learning-Powered Financial Fraud Detection: Building Robust Predictive Models for Transactional Security. Journal of Economics, Finance and Accounting Studies, 5(5), 161–180.

[21] Kadam, P., et al. (2024). Enhancing Financial Fraud Detection with Human-in-the-Loop Feedback and Feedback Propagation. arXiv preprint arXiv:2411.05859.

[22] Khan, M. A. U. H., Islam, M. D., Ahmed, I., Rabbi, M. M. K., Anonna, F. R., Zeeshan, M. D., … & Sadnan, G. M. (2025). Secure Energy Transactions Using Blockchain Leveraging AI for Fraud Detection and Energy Market Stability. arXiv preprint arXiv:2506.19870.

[23] Li, C., Chen, S., Wang, J., & Huang, J. (2021). Reinforcement Learning for Adaptive Fraud Detection in Streaming Environments. IEEE Transactions on Neural Networks and Learning Systems, 32(6), 2485–2496.

[24] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[25] Mahabub, S., Das, B. C., & Hossain, M. R. (2024). Advancing Healthcare Transformation: AI-Driven Precision Medicine and Scalable Innovations through Data Analytics. Edelweiss Applied Science and Technology, 8(6), 8322–8332.

[26] Mahabub, S., Jahan, I., Islam, M. N., & Das, B. C. (2024). The Impact of Wearable Technology on Health Monitoring: A Data-Driven Analysis with Real-World Case Studies and Innovations. Journal of Electrical Systems, 20.

[27] Pelosi, D., Cacciagrano, D., & Piangerelli, M. (2025). Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review. Algorithms, 18(7), 443.

[28] Rahman, M. S., Hossain, M. S., Rahman, M. K., Islam, M. R., Sumon, M. F. I., Siam, M. A., & Debnath, P. (2025). Enhancing Supply Chain Transparency with Blockchain: A Data-Driven Analysis of Distributed Ledger Applications. Journal of Business and Management Studies, 7(3), 59–77.

[29] Rana, M. S., Chouksey, A., Hossain, S., Sumsuzoha, M., Bhowmik, P. K., Hossain, M., … & Zeeshan, M. A. F. (2025). AI-Driven Predictive Modeling for Banking Customer Churn: Insights for the US Financial Sector. Journal of Ecohumanism, 4(1), 3478–3497.

[30] Ray, R. K., Sumsuzoha, M., Faisal, M. H., Chowdhury, S. S., Rahman, Z., Hossain, E., … & Rahman, M. S. (2025). Harnessing Machine Learning and AI to Analyze the Impact of Digital Finance on Urban Economic Resilience in the USA. Journal of Ecohumanism, 4(2), 1417–1442.

[31] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144.

[32] Settles, B. (2009). Active Learning Literature Survey. University of Wisconsin, Madison.

[33] Shovon, M. S. S., Gomes, C. A., Reza, S. A., Bhowmik, P. K., Gomes, C. A. H., Jakir, T., … & Hasan, M. S. (2025). Forecasting Renewable Energy Trends in the USA: An AI-Driven Analysis of Electricity Production by Source. Journal of Ecohumanism, 4(3), 322–345.

[34] Sultana, K. S., Begum, M., Abed, J., Siam, M. A., Sadnan, G. A., Shatyi, S. S., & Billah, M. (2025). Blockchain-Based Green Edge Computing: Optimizing Energy Efficiency with Decentralized AI Frameworks. Journal of Computer Science and Technology Studies, 7(1), 386–408.

[35] Wang, L., Li, Y., & Yuan, F. (2022). Dynamic Ensemble Selection for Streaming Data with Concept Drift. Knowledge-Based Systems, 255, 109685.