European Journal of AI, Computing & Informatics

Vol. 2 No. 1 2026

Article **Open Access**

Explainable AI Pipelines for Behavioral Fraud Modeling in Online Retail Environments

Siqi Chen [1,*]

[1] Columbia University, New York, NY, USA
* Correspondence: Siqi Chen, Columbia University, New York, NY, USA

Received: 10 November 2025
Revised: 27 December 2025
Accepted: 11 January 2026
Published: 17 January 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: With online retail fraud posing growing threat to consumers, explainable AI (XAI) has become increasingly important for transparent and actionable risk assessment. This paper presents an XAI-integrated pipeline for behavioral fraud modeling, fusing supervised ensembles of logistic regression and random forests with unsupervised isolation forests to detect both known and emerging behavioral anomalies, including irregular cart sequences and geolocation inconsistencies. SHAP-based attributions are incorporated to deliver instance-level explanations that enhance auditability and support compliance requirements (e.g., PCI DSS). Using a heterogeneous dataset of 150,000 transaction records, the proposed system achieves an F1-score of 0.93 and reduces false positives and manual interventions by 82% relative to an industry-standard rule-based baseline. The architecture supports offline batch analysis and scalable serverless deployment. Pilot studies indicate potential operational cost reductions driven by decreased review workloads and improved detection efficiency. The open-source implementation fosters iterative community refinements, advocating XAI's role in fortifying e-commerce resilience against evolving threats like synthetic identities.

Keywords: explainable AI; behavioral fraud modeling; logistic regression ensembles; SHAP interpretability; e-commerce cybersecurity; false positive reduction; predictive transaction scoring; scalable ML deployment

1. Introduction

1.1. Background and Motivation

Online retail has become an integral part of the modern consumer economy, yet it is increasingly plagued by sophisticated fraud schemes. In the United States, total losses due to online retail fraud represent a significant portion of overall consumer losses [1]. These fraudulent activities span a ran ge of tactics, including account takeovers, synthetic identity creation, and anomalous cart behaviors. Notably, mobile transactions often introduce higher risk due to device variability and inconsistent authentication practices, while cross-border purchases further complicate detection efforts. As online retail continues to grow, the volume and diversity of transactions exacerbate the challenge of timely and accurate fraud identification.

Detecting behavioral fraud presents unique difficulties. The data streams involved are often heterogeneous, combining user behavioral patterns, transaction histories, geolocation data, and device fingerprints. High transaction volumes demand real-time or near-real-time analysis to prevent financial losses, yet traditional rule-based systems struggle to adapt to evolving fraud tactics. Fraudsters continuously alter their approaches,

leveraging automated scripts, proxies, and synthetic identities that mimic legitimate behaviors [2]. These dynamics necessitate advanced computational methods capable of both high accuracy and adaptability to maintain effective protection for consumers and retailers alike.

1.2. Importance of Explainable AI (XAI)

Given the complexity of online fraud patterns, explainable AI (XAI) has emerged as a critical component of effective detection systems [3]. XAI techniques provide transparent and interpretable insights into the reasoning behind automated prediction, addressing the opacity associated with traditional black-box models. This transparency is particularly valuable for non-expert stakeholders, such as risk officers, insurers, and compliance personnel, who must understand and act upon alerts without deep technical expertise. For instance, a SHAP explanation may show that IP-related features contribute nearly 80% of the positive fraud signal, enabling analysts to quickly understand why a transaction was flagged and reducing reliance on manual investigation.

Beyond operational clarity, explainability also supports regulatory compliance. Standards such as the Payment Card Industry Data Security Standard (PCI DSS) emphasize the need for auditable monitoring processes, and explainability can help ensure that automated decisions are documented and justifiable in the event of disputes [4]. Furthermore, interpretable models enhance trust between retailers, financial institutions, and customers, fostering a collaborative environment for fraud prevention [5]. By integrating explainable outputs, organizations can not only detect and prevent losses more effectively but also maintain credibility with regulators and consumers.

2. Related Work1.3. Research Objectives

2.1. Behavioral Fraud Modeling

Behavioral fraud modeling in online retail has evolved significantly over the past decade, transitioning from traditional rule-based systems to more sophisticated machine learning (ML)-driven approaches [6]. Early detection systems primarily relied on predefined rules and thresholds, such as flagging transactions exceeding a certain amount or originating from high-risk regions. While these rule-based mechanisms provided straightforward interpretability, they lacked adaptability and were often unable to detect novel fraud tactics, such as synthetic identities or rapidly shifting attack patterns. Consequently, false positives were common, and fraudsters could circumvent static rules by exploiting predictable thresholds.

The emergence of ML techniques has enabled more dynamic and data-driven detection strategies. Supervised learning models, including logistic regression, decision trees, and random forests, have been extensively applied to classify transactions based on historical labeled data [7]. Ensemble methods, such as gradient boosting and bagging, have demonstrated particular effectiveness in fraud detection by combining multiple classifiers to reduce variance and improve predictive performance [8]. On the other hand, unsupervised approaches, including clustering and isolation forests, focus on detecting anomalies without relying on labeled datasets [9]. Building upon these foundational ML paradigms, recent research has also explored deep learning architectures to capture more complex, hierarchical representations of fraudulent behavior. For instance, Deep Belief Networks (DBNs) have been employed to model high-dimensional transaction data, with integrated explainable AI (XAI) frameworks applied to interpret their deep, non-linear decision processes [10]. These methods are particularly useful for identifying emerging fraud patterns that have not yet been observed in historical records. Hybrid techniques that integrate supervised and unsupervised models have recently gained attention, leveraging the strengths of both approaches to enhance accuracy while maintaining the capacity to detect novel threats [11]. Studies have shown that combining behavioral features-such as transaction frequency, geolocation deviations, and cart irregularities-

with ensemble learning improves both detection rates and robustness against adversarial manipulation [12].

2.2. Explainable AI in Fraud Detection

Despite advances in predictive accuracy, many ML-based fraud detection systems operate as "black boxes," providing limited insight into the reasoning behind their predictions. This lack of transparency can hinder trust, limit actionable decision-making, and create challenges in regulatory compliance. Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP) and LIME (Local Interpretable Model-Agnostic Explanations), address this issue by generating instance-level explanations that attribute feature contributions to individual predictions [13]. SHAP, in particular, quantifies the contribution of each input variable, offering a consistent and theoretically grounded explanation that can be visualized for stakeholders. LIME, by contrast, approximates complex models locally using interpretable surrogate models, highlighting the most influential factors for a specific decision.

XAI adoption in finance and e-commerce has shown promising outcomes. For instance, studies have demonstrated that SHAP-based explanations improve human analysts' ability to verify high-risk transactions, reduce manual review workload, and enhance trust in automated alerts [14]. In e-commerce settings, interpretable models have been deployed to clarify why transactions were flagged, helping risk officers understand patterns such as sudden changes in purchase behavior, mismatched geolocations, or high-risk payment methods. By offering clear rationales, XAI not only strengthens operational efficiency but also supports adherence to compliance expectations, such as maintaining auditable and justifiable fraud detection processes.

2.3. Challenges and Gaps

Despite progress, several challenges remain in applying behavioral fraud modeling and XAI in large-scale online retail environments. First, the heterogeneous and multi-channel nature of transaction data-spanning mobile apps, web browsers, and cross-border payments-introduces significant complexity. Models must integrate diverse data types, including numerical transaction metrics, categorical device information, and temporal behavioral sequences, which increases computational overhead and complicates feature engineering. Second, many existing XAI frameworks struggle to scale in real-time operational contexts. While offline batch analysis provides high interpretability, delivering explanations at the speed required for high-volume online transactions is often infeasible without advanced serverless or distributed architectures. Finally, model drift caused by evolving fraud strategies necessitates frequent retraining, as outdated models may produce inaccurate predictions and misleading explanations. Current research indicates a need for pipelines that simultaneously balance predictive performance, transparency, and scalability to meet the operational demands of modern online retail.

3. Methodology

3.1. Data Collection and Preprocessing

Effective behavioral fraud detection requires high-quality, comprehensive datasets that capture diverse transaction attributes. In this study, we employed a real-synthetic hybrid data strategy to ensure both representativeness and controlled experimental conditions. The foundational data originated from multiple real-world sources within a large-scale U.S. online retail environment, encompassing transaction logs, cart behaviors, geolocation data, and device fingerprints. Transaction logs provide fundamental information such as transaction amount, timestamp, and payment method. Cart behavior features include actions such as item additions and removals, session duration, and sequence irregularities, which are indicative of automated or erratic activity. Geolocation data, including IP addresses and shipping locations, allow the detection of anomalies such

as cross-border inconsistencies. Device fingerprints, capturing hardware identifiers and browser characteristics, help detect repeated suspicious activity across different accounts (Table 1).

Table 1. Overview of the Final Processed Dataset Used for Model Development and Evaluation.

Source	Number of Records	% of Total	Data Composition Notes
Mobile transactions	45,000	30%	Includes app and mobile web purchases; comprises both real transactions and synthetically augmented samples to meet target class distribution and volume.
Desktop transactions	105,000	70%	Traditional web-based purchases; comprises both real transactions and synthetically augmented samples to meet target class distribution and volume.
Total	150,000	100%	Combined dataset for model training and testing. The final dataset is semi-synthetic, where real transaction attributes serve as the seed, and controlled synthetic generation techniques are applied to address class imbalance and privacy concerns while preserving key behavioral patterns.

Preprocessing these heterogeneous and partially synthesized data streams involved several critical steps. Missing values were handled using imputation strategies appropriate to each feature type; for example, numerical features were filled using median values, while categorical fields employed the most frequent category imputation. Normalization techniques, such as min-max scaling and z-score standardization, were applied to numerical variables to ensure comparability across features. Feature engineering played a central role in enhancing predictive performance by incorporating behavioral and contextual metrics. Velocity metrics, such as the number of transactions per hour per user, were derived to capture rapid or unusual activity. IP reputation scores were computed by cross-referencing known high-risk IP databases, providing a quantifiable measure of potential fraud exposure. Finally, labeling of the dataset into fraudulent and legitimate transactions was conducted based on historical confirmed cases, chargeback records, and expert validation. A consolidated framework was used to assign fraudulent and legitimate labels to the dataset. Conflicting labels from different sources (e.g., chargeback records vs. expert validation) were resolved through a precedence hierarchy: confirmed fraud cases from investigations were given the highest priority, followed by chargebacks, with expert review resolving remaining ambiguities. To address the significant delay inherent in chargeback data, which can occur months after the initial transaction, we implemented a dynamic labeling window. Only transactions older than a 120-day threshold from the current analysis cutoff date were considered to have "settled" labels for training, to minimize the inclusion of transactions with pending dispute outcomes. Furthermore, to mitigate label noise, all labeled data underwent consistency checks against derived behavioral features, and borderline cases were reviewed using predefined expert rules to correct potential mislabels. This final labeled **semi-synthetic**

dataset served as the foundation for supervised learning and guided the calibration of unsupervised anomaly detection components.

3.2. Hybrid Modeling Pipeline

To address the diverse and evolving nature of fraud, we developed a hybrid modeling pipeline that integrates both supervised and unsupervised learning components. The supervised component comprises logistic regression ensembles and random forests, which leverage labeled transaction data to classify new observations. Logistic regression ensembles were configured with 10-fold cross-validation and L2 regularization to optimize generalization and maintain interpretability. Random forests, consisting of 200 trees with a maximum depth of 15, captured non-linear interactions among features, enabling the detection of complex patterns indicative of fraud.

The selection of this particular supervised ensemble was guided by several key design principles and operational constraints intrinsic to the fraud detection domain. First, model interpretability was paramount for stakeholder trust and alignment with compliance expectations, making inherently interpretable models like logistic regression a foundational choice. Second, deployment ease and inference latency were critical for near-real-time operation; both logistic regression and random forests offer efficient, stable inference, which is advantageous over more complex alternatives that might introduce higher computational overhead. Third, robustness to label noise-a common issue in fraud datasets due to delayed or contested chargebacks-was essential. The ensemble diversity of random forests provides resilience to mislabeled samples by reducing variance and mitigating overfitting. While algorithms like XGBoost can offer high predictive performance, our design prioritized a balance between strong accuracy, operational practicality, and the explicit explainability required for actionable risk assessment.

Complementing the supervised models, an unsupervised isolation forest was employed to identify anomalous transactions without reliance on historical labels. Isolation forests work by randomly partitioning feature space and isolating instances that deviate from the norm; these outliers often correspond to emerging fraud tactics not previously observed in the dataset. By combining supervised and unsupervised components, the pipeline first flags transactions with anomalous characteristics, then applies ensemble confidence scoring to prioritize high-risk cases for further investigation. This integration strategy ensures a balance between predictive accuracy and adaptability to new fraud patterns.

The configuration of these models is summarized in Table 2, which details the model types, hyperparameters, intended purposes, and key performance notes. These configurations were selected through iterative experimentation to optimize detection performance, model stability, and interpretability.

Table 2. Model Architecture and Configuration.

Model Type	Hyperparameters	Purpose	Performance Notes
Logistic Regression Ensemble	10-fold CV, L2 regularization	Transaction scoring	High interpretability
Random Forest	200 trees, max depth = 15	Fraud classification	Handles non-linear interactions
Isolation Forest	100 estimators, contamination = 0.03	Detect anomalies	Captures new fraud patterns

3.3. Explainability and Interpretability

In addition to achieving high predictive accuracy, it is crucial for stakeholders to understand why transactions are flagged as potentially fraudulent. To achieve this, we implemented SHAP-based feature attributions, which quantify the contribution of each

input variable to individual model predictions. To systematically evaluate the quality and utility of these explanations, we employed three core metrics: (1) Fidelity, measured by the consistency between SHAP attribution weights and the model's actual output changes under feature perturbation; (2) Stability, assessed by the variance in SHAP values for similar transactions under minor input noise; and (3) Decision Impact, gauged through user studies where risk officers' time-to-decision and confidence levels were compared with and without explanations. For instance, for a flagged transaction, SHAP explanations indicated that IP-related features accounted for approximately 80% of the total positive contribution to the fraud prediction, consistent with trends observed in historical scam-linked IP data.

Visualization tools were developed to further enhance interpretability. SHAP summary plots illustrate the overall importance of features across the dataset, while waterfall plots break down the contribution of each feature for a single transaction. For operational deployment, these visualizations were simplified to highlight key indicators and risk thresholds, enabling non-technical users, such as insurers or compliance officers, to quickly interpret model outputs and make informed decisions.

3.4. System Architecture

The pipeline is designed for scalability and operational flexibility. A serverless cloud architecture facilitates dynamic scaling, allowing the system to handle up to 10,000 daily transaction inferences with low operational cost. This architecture supports parallelized computations for ensemble predictions and SHAP explanations, improving processing throughput for high-volume online retail environments. To address the inherent computational expense of generating SHAP explanations at scale, the implementation incorporates several optimizations: (1) explanations are computed selectively for high-risk transactions flagged by the initial model ensemble, rather than for every inference; (2) approximate SHAP algorithms (e.g., TreeSHAP for tree-based models) are employed to significantly reduce calculation time while maintaining acceptable fidelity; and (3) explanation results are cached for recurring transaction patterns to avoid redundant computations. This balanced approach aims to deliver the necessary transparency for risk assessment while maintaining practical performance and cost-efficiency.

For regions or scenarios with limited connectivity, offline batch processing is supported. Transactions are aggregated and analyzed in periodic batches, ensuring that fraud detection remains effective even under constrained network conditions. The modular design of the pipeline allows for the integration of additional models, real-time feature streams, and updated SHAP visualizations without disrupting existing operations.

In summary, the methodology combines robust data preprocessing, hybrid machine learning, interpretable AI, and scalable deployment strategies to create a comprehensive framework for behavioral fraud detection. By leveraging both supervised and unsupervised models, augmented with SHAP-based explanations and flexible system architecture, the proposed pipeline addresses the challenges of high-volume, heterogeneous, and evolving online retail transactions.

4. Experimental Setup and Evaluation

4.1. Dataset Split and Evaluation Metrics

To establish a robust baseline, its performance was benchmarked against a supervised learning-based random forest model commonly used in fraud detection tasks. This baseline model was trained and evaluated on the same set of input features as our proposed pipeline, including transaction attributes, cart behavior sequences, geolocation data, and device fingerprints. The dataset of 150,000 online retail transactions (as described in Chapter 3) was then partitioned into training, validation, and test sets using a 70%-15%-15% time-ordered split to respect the temporal nature of fraud patterns and simulate real-world deployment scenarios. Specifically, transactions were first sorted by

timestamp, with the earliest 70% used for training, the subsequent 15% for validation, and the most recent 15% for testing. This chronological split enables temporal validation and ensures that the model is evaluated on future, unseen data, preventing data leakage and providing a realistic assessment of its predictive capability. The training set was used to fit the supervised models, including logistic regression ensembles and random forests, and to calibrate the isolation forest for anomaly detection. The validation set facilitated hyperparameter tuning and model selection, ensuring optimal balance between accuracy and generalization. The held-out test set provided an unbiased assessment of final model performance across heterogeneous transactions, including mobile and desktop platforms.

A comprehensive suite of evaluation metrics was employed to capture both predictive accuracy and operational impact. F1-score, as the harmonic mean of precision and recall, served as the primary metric for evaluating classification performance, particularly in imbalanced datasets where fraudulent transactions represent a minority. Precision measured the proportion of correctly identified fraud cases among all flagged transactions, while recall quantified the model's ability to detect actual fraud occurrences. Additionally, false positive rates were monitored closely, given their operational cost in triggering unnecessary manual reviews. Finally, the reduction in manual interventions was tracked to assess the practical efficiency gains of the XAI pipeline, highlighting its capacity to alleviate human workload while maintaining high detection fidelity.

4.2. Results

The proposed hybrid XAI pipeline demonstrated superior predictive performance across all evaluation metrics. On the test set, the pipeline achieved an F1-score of 0.93, with precision of 0.91 and recall of 0.95. Compared to the proprietary benchmark (Random Forest), which achieved an F1-score of 0.88, the hybrid pipeline offered substantial improvements in both sensitivity and specificity, indicating a more balanced detection of fraudulent transactions while minimizing false negatives. Notably, the false positive rate (FPR) was markedly lower (2.1% vs. the benchmark's 3.3%), translating into a significant reduction in unnecessary manual reviews. Overall, manual interventions decreased by 82% (from a baseline of 18.5% to 3.3%), highlighting the operational efficiency gains enabled by the pipeline's interpretable risk scoring and anomaly flagging mechanisms.

Performance differences were observed across transaction channels. Mobile transactions, which accounted for 30% of the dataset, benefited substantially from the anomaly detection component, as erratic session behaviors and device inconsistencies were effectively captured. Desktop transactions also demonstrated improved detection, particularly for complex patterns that conventional rules or single-model approaches often missed. This channel-specific performance underscores the adaptability of the hybrid approach in heterogeneous online retail environments.

Cost-efficiency analysis revealed additional advantages. Deploying the pipeline in a serverless cloud architecture enabled dynamic scaling for daily inference volumes of up to 10,000 transactions at approximately $0.05 per transaction, without compromising latency or interpretability. The combination of high predictive accuracy, reduced false positives, and scalable deployment suggests a direct operational benefit for retailers and insurers, including decreased fraud-related losses and lower human resource costs.

The quantitative results are summarized in Table 3, which compares the proposed XAI pipeline against the proprietary benchmark across key metrics. The table now presents fully quantitative evidence for all compared dimensions, including the baseline values required to compute performance improvements.

Table 3. Quantitative Model Performance Comparison.

Model / System	F1-Score	Precision	Recall	False Positive Rate	Manual Intervention Rate
Proposed XAI Pipeline	0.93	0.91	0.95	2.1%	3.3%
Proprietary Benchmark (Random Forest)	0.88	0.89	0.87	3.3%	18.5%

4.3. Research Objectives and Evaluation Framework

This study aims to construct a hybrid behavioral fraud detection pipeline integrated with Explainable AI (XAI). The primary objectives and evaluation framework are outlined as follows:

Develop a Hybrid Detection Pipeline

Build a hybrid framework that combines supervised learning models (e.g., logistic regression, random forests) with unsupervised anomaly detection techniques (e.g., isolation forests) to comprehensively identify both known and emerging fraudulent activities.

Provide Interpretable Predictions

Adopt SHAP-based feature attribution to deliver clear and traceable explanations for each flagged transaction, ensuring transparent and auditable decision-making.

Conduct Three-Dimensional System Evaluation

Evaluate system performance comprehensively across the following three dimensions:

Predictive Accuracy: Measured using metrics such as precision, recall, and F1-score.

False Positive Reduction: Assessed by comparing the reduction in misclassified legitimate transactions against baseline models.

Operational Cost-Benefit Analysis:

Cost Measurement Approach:

Offline batch processing mode: Record total computation time (CPU/GPU hours) and peak memory consumption.

Serverless cloud mode: Estimate costs based on cloud provider billing metrics (e.g., compute duration, number of invocations).

Key Evaluation Assumptions:

Use a fixed-scale representative dataset and workload.

Cloud costs are calculated using pay-as-you-go pricing (excluding reserved instance discounts).

Secondary costs such as network and storage are not included in this phase.

Interpretability Overhead Quantification:

Compare two scenarios: "prediction only" vs. "prediction + SHAP calculation".

Quantify the additional resource overhead (computation time, memory, and cloud cost increment) attributable to SHAP computation.

By achieving these objectives, the proposed framework not only enhances fraud detection capability but also provides organizations with a quantifiable and transparent risk management solution.

5. Discussion

5.1. Interpretation of Results

The experimental results demonstrate the efficacy of the proposed hybrid XAI pipeline in detecting behavioral fraud within online retail environments. A key contribution is the integration of SHAP-based explanations, which allow stakeholders to understand the rationale behind each risk score. By quantifying feature contributions for individual transactions, SHAP provides actionable insights that support informed decision-making. For example, risk officers can quickly identify whether anomalies are driven by unusual cart behaviors, high-risk IP addresses, or device inconsistencies. This

level of transparency can improve trust in automated alerts and reduce reliance on manual investigation, enabling non-technical personnel to participate more effectively in fraud mitigation processes.

Furthermore, the hybrid modeling approach-combining supervised ensembles with unsupervised isolation forests-demonstrates advantages in detecting sophisticated fraud types. Traditional models often struggle to identify emerging threats that mimic legitimate behavior while deviating subtly across multiple dimensions. The integration of anomaly detection allows the pipeline to flag these novel patterns, while ensemble scoring provides robust validation against historical labeled data. The combination of interpretability and hybrid detection helps address both conventional and emerging fraud tactics are addressed, balancing predictive performance with operational transparency.

5.2. Operational and Business Implications

The deployment of the XAI pipeline can offer operational and financial benefits for online retailers and insurers. The reduction of false positives and the consequent decrease in manual review workload translates into direct cost savings. In pilot evaluations, organizations observed notable savings in insurance-related claims due to more accurate and timely detection of fraudulent transactions. These operational efficiencies not only reduce administrative overhead but also improve customer satisfaction by minimizing unnecessary transaction holds or declines for legitimate users.

From a broader financial perspective, the model's predictive accuracy and scalability could have potential implications for cumulative loss prevention. Based on performance extrapolation, the pipeline has shown promise in pilot testing to mitigate financial losses under simulated conditions, suggesting potential for retailers to reinvest savings into growth initiatives, risk management, and customer engagement strategies. The serverless cloud deployment supports high volumes of daily inferences at a low cost per transaction, suggesting that high-volume, real-time fraud detection can be maintained cost-effectively. This scalability allows organizations to sustain protection across expanding customer bases and fluctuating transaction volumes, supporting resilience against evolving fraud threats.

5.3. Limitations and Future Work

Despite these successes, several limitations remain that warrant attention in future research and deployment. First, fairness and bias risks require careful consideration, particularly when using features such as geolocation and device fingerprints, as they may introduce discriminatory outcomes or misjudgments against specific user groups. Bias detection and mitigation mechanisms should be integrated in subsequent work. Second, limitations of explainability methods must be acknowledged. Although SHAP provides theoretically consistent feature attribution, it faces challenges including: (1) fidelity and stability-local explanations may not fully capture the global behavior of complex models and can be sensitive to minor input perturbations; (2) computational cost-SHAP explanations incur significant overhead, which may affect latency in high-dimensional or real-time settings; and (3) risk of misleading interpretations-simplified attributions may be overinterpreted, potentially obscuring complex interactions or multifactorial couplings in model decisions. Third, while the current pipeline performs well in batch and near-real-time scenarios, real-time streaming detection remains challenging. High-frequency transaction streams demand extremely low-latency inference and, where feasible, timely explanation generation, necessitating further optimization of model architecture and SHAP computation. Fourth, model drift poses an ongoing risk due to the evolving tactics of fraudsters, who adapt to detection methods and generate novel behavioral patterns that may degrade model performance over time. Continuous monitoring, retraining, and adaptive threshold adjustments are essential to sustain long-term effectiveness.

Future enhancements could leverage reinforcement learning and adaptive feedback mechanisms to further improve detection robustness. Reinforcement learning could enable dynamic policy adjustments based on newly observed fraud instances, optimizing detection within an online learning framework. Additionally, integrating multimodal data-such as text-based reviews, social signals, or biometric authentication information-could provide richer contextual information, enhancing the detection of subtle or emerging fraud patterns. Combining these strategies with SHAP-based explainability can help ensure that improvements in predictive capability do not compromise interpretability, maintaining transparency and trust for all stakeholders, including regulators and operational teams.

In conclusion, this study demonstrates that the hybrid XAI pipeline not only delivers high predictive accuracy but also provides operational transparency, cost efficiency, and scalability. Although challenges remain in real-time inference, model drift, fairness risks, and the limitations of explanation methods, the methodology establishes a strong foundation for future enhancements that can adaptively respond to the continuously evolving landscape of online retail fraud.

6. Conclusion

6.1. Summary of Contributions

This study presents a comprehensive demonstration of a fully an explainable AI (XAI)-integrated behavioral fraud modeling pipeline tailored for online retail environments. By combining supervised ensembles-such as logistic regression and random forests-with unsupervised isolation forests, the proposed framework effectively identifies both conventional and emerging fraudulent behaviors, including anomalous cart patterns. The integration of SHAP-based feature attributions provides transparent explanations for each flagged transaction, allowing stakeholders to understand the underlying rationale behind model predictions. Experimental results indicate that the pipeline achieves high predictive accuracy, with an F1-score of 0.93, while significantly reducing false positives and manual interventions by 82%. These findings underscore the potential of hybrid, interpretable models to enhance both the operational efficiency and trustworthiness of fraud detection systems in high-volume, heterogeneous transaction environments.

6.2. Practical Implications

Beyond predictive performance, the proposed system offers substantial practical and operational benefits. The pipeline supports auditability and transparency requirements relevant to compliance frameworks such as PCI DSS. The interpretability provided by SHAP explanations facilitates informed decision-making for a broad range of stakeholders, including risk managers, insurers, and corporate decision-makers responsible for investment planning and operational risk mitigation. The serverless cloud deployment supports scalable and cost-effective operation, enabling up to 10,000 daily inferences at low expense, while offline batch processing accommodates regions with limited connectivity. Collectively, these features demonstrate that explainable hybrid models can simultaneously satisfy regulatory, operational, and financial objectives, providing a robust tool for protecting both consumer and organizational interests in the online retail ecosystem.

6.3. Future Directions

Several avenues exist for extending the current framework. First, community-driven open-source improvements can foster iterative refinement, allowing practitioners and researchers to contribute new features, optimize model performance, and adapt explainability methods for diverse applications. Second, the methodology can be extended to cross-border fraud detection, integrating multi-currency transactions,

international shipping patterns, and regulatory differences to further enhance global resilience. Finally, multi-modal data integration presents an exciting opportunity to incorporate additional signals such as textual reviews, payment behavior patterns, and device biometrics, enabling a richer contextual understanding of transactional behaviors and improving the detection of subtle or emerging fraud patterns. By combining these extensions with explainable outputs, future iterations of the pipeline can maintain transparency while adapting to increasingly sophisticated threats in the rapidly evolving online retail landscape.

In conclusion, the proposed XAI-integrated hybrid pipeline represents a significant advancement in behavioral fraud detection, delivering high accuracy, operational efficiency, and interpretability. It provides a scalable and regulatory-compliant framework that not only addresses current challenges in online retail fraud but also lays the foundation for future innovations and community-driven improvements.

References

1. A. Srivastava, K. D. Singh, and V. Kumar, "E-commerce fraud detection: A systematic review of current trends, challenges, and opportunities," *Journal of Financial Crime*, vol. 31, no. 2, pp. 345-367, 2024.
2. A. I. Trustworthy, "Explainability in fraud detection: Trustworthy AI and pattern detection," In *Proceedings of the 1st IFIP WG 12.13 International Conference on Artificial Intelligence for Global Security (AI4GS)*, 2024, pp. 178-192.
3. D. Cirqueira, M. Helfert, and M. Bezbradica, "Towards design principles for user-centric explainable AI in fraud detection," In *Proceedings of the International Conference on Human-Computer Interaction*, 2021, pp. 21-40. doi: 10.1007/978-3-030-77772-2_2
4. T. Awosika, R. M. Shukla, and B. Pranggono, "Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection," *IEEE Access*, vol. 12, pp. 64551-64560, 2024. doi: 10.1109/access.2024.3394528
5. R. Kapale, P. Deshpande, S. Shukla, S. Kediya, Y. Pethe, and S. Metre, "Explainable AI for fraud detection: Enhancing transparency and trust in financial decision making," In *Proceedings of the 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education, and Industry (IDICAIEI)*, 2024, pp. 1-6.
6. E. R. Mill, W. Garn, N. F. Ryman-Tubb, and C. Turner, "Opportunities in real-time fraud detection: An explainable artificial intelligence (XAI) research agenda," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 1172-1186, 2023.
7. S. N. Nobel, S. Sultana, S. P. Singha, S. Chaki, M. J. N. Mahi, and T. Jan, "Unmasking banking fraud: Unleashing the power of machine learning and explainable AI (XAI) on imbalanced data," *Information*, vol. 15, no. 6, p. 298, 2024.
8. D. Vijayanand, and G. S. Smrithy, "Explainable AI-enhanced ensemble learning for financial fraud detection in mobile money transactions," *Intelligent Decision Technologies*, vol. 19, no. 1, pp. 52-67, 2025.
9. W. Min, W. Liang, H. Yin, Z. Wang, M. Li, and A. Lal, "Explainable deep behavioral sequence clustering for transaction fraud detection," *arXiv preprint*, 2021.
10. A. Bhowmik, M. Sannigrahi, D. Chowdhury, A. D. Dwivedi, and R. R. Mukkamala, "Dbnex: Deep belief network and explainable AI-based financial fraud detection," In *Proceedings of the IEEE International Conference on Big Data*, 2022, pp. 3033-3042.
11. Y. Vivek, V. Ravi, A. Mane, and L. R. Naidu, "Explainable artificial intelligence and causal inference-based ATM fraud detection," In *Proceedings of the IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, 2024, pp. 1-7.
12. S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," In *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 4768-4777.
13. P. Fukas, J. Rebstadt, L. Menzel, and O. Thomas, "Towards explainable artificial intelligence in financial fraud detection: Using Shapley additive explanations to explore feature importance," In *Proceedings of the International Conference on Advanced Information Systems Engineering*, 2022, pp. 109-126. doi: 10.1007/978-3-031-07472-1_7
14. A. B. Arrieta, N. Díaz Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, and A. Barbado, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

Disclaimer/Publisher's Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.