International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 288-295, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P148

Original Article

Behavioral Pattern Analysis for Return Fraud Detection in High-Volume E-Commerce: A Multi-Signal Approach

Deepanjan Mukherjee
Independent Researcher, Austin, TX USA.

Received On: 31/01/2026 ***Revised On: 02/03/2026*** ***Accepted On: 04/03/2026*** ***Published on: 08/03/2026***

Abstract – *Return fraud costs U.S. retailers $103 billion annually, yet academic fraud detection literature remains almost entirely focused on payment transaction fraud. This paper presents a multi-signal behavioral analysis framework for return fraud detection that fuses three complementary signal streams: transaction-level features, behavioral history profiles, and natural language processing of customer-supplied return reason text. A late-fusion stacking architecture combines independently trained branch classifiers through a gradient-boosted meta-learner, enabling modular updates as fraud patterns evolve. Cost-sensitive learning, calibrated to the financial value of individual transactions, ensures that the system's optimization objective aligns with actual business impact rather than raw classification accuracy. A five-component evaluation methodology employs temporal validation and example-dependent cost metrics to establish realistic performance benchmarks that account for concept drift and class imbalance inherent to high-volume retail environments. The proposed system targets greater than 90% sensitivity for high-severity fraud categories while maintaining greater than 85% specificity to protect legitimate customer relationships. By integrating NLP analysis of return justification text a signal channel absent from all existing fraud detection literature despite a documented 76% consumer embellishment rate this framework addresses a structural gap in both academic research and commercial practice.*

Keywords – *Return Fraud Detection, E-Commerce Fraud, Behavioral Anomaly Detection, Natural Language Processing, Multi-Signal Fusion, Risk Scoring, Cost-Sensitive Learning, Concept Drift, Imbalanced Classification, AI-Assisted Fraud Detection.*

1. Introduction

The consumer return process has become one of the largest unmanaged cost centers in modern retail. In 2024, U.S. retailers processed $685 billion in returns, of which $103 billion (15.14%) were attributable to fraud and abuse [1]. These figures represent a meaningful increase from $101 billion in fraud losses against $743 billion in total returns recorded the prior year [2], and projections for 2025 place total returns at $849.9 billion with fraud continuing to grow as a proportion [3]. The scale of this problem exceeds the annual revenue of many Fortune 500 companies, yet it has received almost no systematic academic attention.

Return fraud takes several distinct forms, each exploiting a different vulnerability in retailer return policies. Wardrobing, the practice of purchasing items for temporary use before returning them, is the most prevalent: 69% of consumers surveyed admit to having wardrobed at least once, with 64% reporting they do so at least monthly [4]. Empty box fraud, where customers return packages containing substituted or missing merchandise, accounts for 31% of all fraudulent returns and is reported as the leading fraud indicator by retailers tracking abuse patterns [1]. Bracketing, purchasing multiple variants of an item with the intent to return all but one, is practiced by 51% of Generation Z consumers [5]. Receipt fraud and cross-retailer return schemes round out the primary fraud taxonomy.

What distinguishes return fraud from the payment fraud that dominates academic literature is the participation of the customer in constructing the fraudulent event. A cardholder whose credentials are stolen is a victim; a customer committing wardrobing is an active agent who interacts with the return system, generates behavioral history, and supplies a textual reason for their return. The 76% of consumers who admit to embellishing return reasons [4], a figure 39% higher than the prior year, create a natural language signal that is both structurally rich and, to date, entirely unexploited in academic fraud detection research.

The machine learning fraud detection literature, which Mutemi and Bacao [6] systematically reviewed across 101 publications, is concentrated almost entirely on payment transaction fraud unauthorized card use, account takeover, and identity theft. Behavioral modeling approaches such as the hierarchical explainable network of Zhu et al. [7] and the competitive graph neural network framework of Zhang et al. [8] operate on purchase transaction sequences and do not model return behavior as a fraud signal. No published work presents a framework specifically designed for return fraud detection, and none incorporates text analysis of customer-provided return justifications.

This paper makes four contributions. First, it provides a systematic characterization of the return fraud problem space, distinguishing it structurally from payment fraud and identifying the signal types that distinguish it. Second, it proposes a three-branch multi-signal framework combining transaction-level features, behavioral history profiles, and NLP analysis of return reason text. Third, it introduces a late-fusion stacking architecture with cost-sensitive learning calibrated to individual transaction value, which aligns detection performance with financial outcomes. Fourth, it proposes a five-component evaluation methodology using temporal validation and example-dependent cost metrics, following the realistic fraud modeling approach of Dal Pozzolo et al. [9] and adapted for the specific label-delay characteristics of return fraud.

The remainder of this paper proceeds as follows. Section II reviews background literature across return fraud, payment fraud detection, NLP-based fraud identification, and the technical foundations of cost-sensitive learning and concept drift. Section III presents the multi-signal framework design. Section IV describes the proposed evaluation methodology. Section V discusses implications and limitations, and Section VI concludes with directions for future work.

2. Background

2.1. Return Fraud: Problem Characteristics and Existing Detection

Return fraud is operationally distinct from payment fraud in three ways that carry direct architectural implications. First, the fraudulent actor is typically an authenticated customer with a legitimate account history, rather than an unauthorized third party. Account-level identity signals that effectively detect stolen credentials are therefore weak discriminators for return abuse. Second, the fraud event spans a longer temporal window: a wardrobing customer may purchase, use, and return items over 14 to 30 days, with no single transaction appearing anomalous in isolation. Third, the ground truth label is often ambiguous and delayed determining whether a return was genuinely fraudulent may require physical inspection of returned merchandise, which introduces labeling latency of days to weeks.

Current detection approaches employed by retailers are largely rule-based: return frequency thresholds, dollar caps, and cross-retailer blacklists maintained by services such as Appriss Retail's Verify-1 platform. These systems compare individual return requests against static thresholds and known-fraud registries. While operationally simple, rule-based approaches suffer from two documented weaknesses. They produce high false positive rates among legitimate customers with unusual but honest return behaviors, and they are trivially circumvented by fraudsters who learn threshold parameters through trial and error. No published academic study has evaluated a learning-based approach to this specific problem, and no public benchmark dataset exists for return fraud modeling.

2.2. Machine Learning for Payment Fraud Detection

The broader fraud detection literature provides technical foundations that are partially transferable to the return fraud domain. Ensemble methods have demonstrated consistent effectiveness on transaction fraud: stacking architectures that combine gradient-boosted trees, random forests, and logistic regression through a meta-learner achieve F1 scores above 88% on public credit card benchmarks [6]. The European credit card fraud dataset (284,807 transactions, 0.172% fraud rate) and the IEEE-CIS Fraud Detection dataset (590,540 transactions) are the two primary public benchmarks, though both cover payment fraud exclusively.

Graph Neural Networks have become the dominant approach for e-commerce fraud at a platform scale. Zhang et al. [8] introduced eFraudCom, which models user-item purchase relationships as a competitive bipartite graph to detect coordinated fraud rings on the Alibaba platform. Dou et al. [10] addressed the camouflage problem, where fraudsters mimic legitimate behavioral patterns, with CARE-GNN, which uses reinforcement-enhanced neighbor sampling to distinguish genuine from disguised edges. Liu et al. [11] developed PC-GNN to handle extreme class imbalance of fraud graphs through a pick-and-choose under-sampling strategy at the graph level. While effective for payment fraud at platform scale, these approaches carry substantial computational overhead and depend on dense relational structures among users a dependency that does not hold for individual retail return fraud, where each fraud event is largely independent.

Tax et al. [12] explicitly identified return fraud as one of the most underserved problems in the e-commerce fraud research agenda, noting the absence of public datasets and systematic feature engineering approaches as primary obstacles to progress. Their work formalizes the research gap this paper directly addresses.

3. Natural Language Processing for Fraud Detection

The application of NLP to fraud detection has proceeded primarily in insurance and financial domains. Yang et al. [13] demonstrated that multimodal fusion of structured claim data with unstructured text outperforms either modality alone for auto insurance fraud, achieving an AUC improvement of 8.3 percentage points over structured-only baselines. Their architecture uses BERT-based embeddings of claim narratives fused with structured tabular features through a dual-channel network, a design principle directly applicable to return reason text.

Taneja et al. [14] applied transformer-based classification to online recruitment fraud, developing Fraud-BERT, a domain-adapted BERT variant, that achieves F1 of 0.93 on the EMSCAD job posting dataset. Their work demonstrates that domain-specific fine-tuning of pre-trained language models substantially outperforms general-purpose text classification on fraud tasks. Saddi et al. [15] surveyed NLP techniques applied to insurance claims across named entity recognition, sentiment analysis, and topic modeling,

finding consistent gains from combining rule-based pattern matching with learned text representations.

The return fraud context presents a distinct NLP challenge. Return reason text is short, typically one to three sentences, written in response to a constrained multiple-choice taxonomy supplemented by an open text field, and exhibits characteristic linguistic markers: vagueness, inconsistency with purchase patterns, and formulaic phrasing. These properties differ structurally from the longer narrative claims analyzed in insurance fraud literature, and no published study has applied any text analysis technique to retail return reason text.

4. Behavioral Sequence Modeling in E-Commerce

Behavioral history modeling for fraud has been explored most extensively by Zhu et al. [7], who applied a hierarchical explainable network to cross-domain fraud detection using purchase event sequences. Their work demonstrates that temporal aggregations of account-level purchase behavior inter-event timing, category concentration, and transaction velocity carry substantial discriminative signal even when individual events appear legitimate in isolation. While Zhu et al. operate exclusively in the payment fraud domain, the same aggregation principle applies directly to return behavior: return frequency, days-to-return variance, and category-level return rates are the return-domain analogues of the purchase-sequence features that prove discriminative in their framework. Zhang et al. [16] analyzed user behavior data from a major e-commerce platform and identified specific behavioral signatures distinguishing fraudulent accounts: purchase-return cycles shorter than 48 hours, category switching inconsistent with account history, and address cycling across orders.

These behavioral approaches operate primarily in the payment fraud domain. Their core insight, that fraud leaves statistical traces in account-level behavioral history that individual events do not exhibit, transfer directly to return fraud. A customer who returns 40% of all purchases, consistently within 48 hours of delivery, and whose return reasons cycle through the same small set of phrases, exhibits a behavioral profile statistically distinguishable from legitimate return behavior even when each individual event appears plausible in isolation.

5. Technical Foundations

Three technical areas underpin the proposed framework. First, cost-sensitive learning: in fraud detection, the financial consequence of a missed fraud event is not uniform missing a $500 fraud incident costs 50 times more than missing a $10 incident. Bahnsen et al. [17] developed a cost-sensitive feature engineering framework for credit card fraud using calendar-based aggregations weighted by transaction amounts and separately demonstrated Bayes minimum risk classification for cost-weighted detection [18]. Their framework formalizes savings, the net financial benefit of detection relative to a no-model baseline, as the primary optimization metric.

Second, concept drift: return fraud patterns shift continuously as fraudsters adapt to detection policies; seasonal assortments change, and macroeconomic conditions alter consumer behavior. Gama et al. [19] provide the foundational taxonomy of drift types sudden, gradual, incremental, and recurring and demonstrate that static models trained on historical data degrade predictably under each type. Detection systems must incorporate mechanisms to identify drift and trigger retraining.

Third, imbalanced classification: fraud events constitute a small fraction of all transactions, typically 1–5% even in high-fraud environments. Chawla et al. [20] introduced SMOTE (Synthetic Minority Over-sampling Technique) as a principled approach to generating synthetic minority-class examples. Fernández et al. [21] surveyed 15 years of progress on SMOTE variants, identifying ensemble-based approaches as consistently superior to single-model resampling. Saito and Rehmsmeier [22] demonstrated that AUC-PR is more informative than AUC-ROC for imbalanced classification an important methodological point for fraud benchmark comparisons.

6. Multi-Signal Framework Design

6.1. Framework Overview

The proposed framework decomposes return fraud detection into three parallel signal branches, each trained independently on feature sets suited to a different data modality, before combining their outputs through a late-fusion stacking meta-learner. Figure 1 illustrates the high-level architecture. The three signal branches are: (1) a Transaction Branch operating on product-level and order-level features from the return event itself; (2) a Behavioral Branch operating on account-level temporal aggregations of return history; and (3) an NLP Branch operating on the textual return reason provided by the customer. Each branch produces a probability estimate and a confidence score. These outputs feed a gradient-boosted meta-learner that produces the final fraud probability, and a severity-weighted risk score calibrated against transaction value.

Late fusion is preferred over early fusion, concatenating all features before training a single model, or intermediate fusion for three reasons. It allows each branch to be retrained independently when drift is detected in that modality's feature distribution, reducing model maintenance cost. It preserves interpretability by attributing predictions to specific signal sources. And it allows the system to operate gracefully when one signal source is unavailable: if a customer provides no free-text return reason, the NLP branch abstains and the meta-learner weights the remaining two branches accordingly.

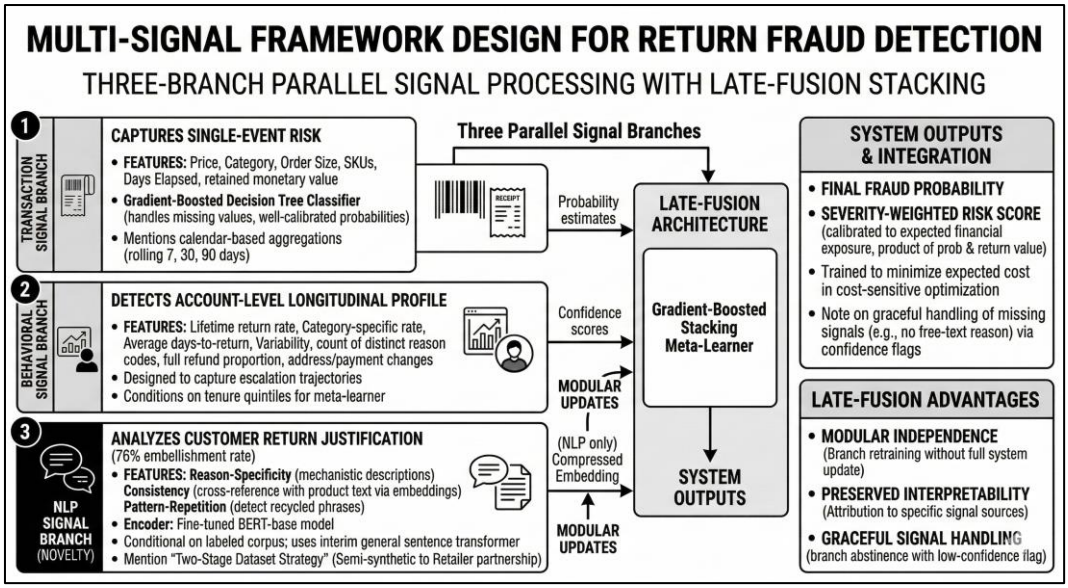

Fig 1: Multi-Signal Framework Design for Return Fraud Detection

6.2. Transaction Signal Branch

The Transaction Branch captures features derived from the return event and its corresponding purchase record. These include item-level features (unit price, product category, brand tier, seasonal classification), order-level features (order size, number of distinct SKUs, shipping method, discount applied), and return-event features (days elapsed between purchase and return, number of items returned relative to items ordered, monetary value of items retained). These features characterize the individual transaction's risk exposure independent of account history.

Feature engineering follows the calendar-based aggregation framework of Bahnsen et al. [17], computing rolling aggregations at 7-day, 30-day, and 90-day windows to capture both acute spikes and chronic patterns. For the Transaction Branch, these aggregations operate at the product-category level, capturing anomalous return rates within specific item classes that may signal product-specific fraud schemes, such as electronics return fraud targeting high-resale-value items. A gradient-boosted decision tree ensemble serves as the branch classifier. These algorithms are robust to heterogeneous feature scales, handle missing values natively, and produce well-calibrated probability estimates essential for downstream fusion.

6.3. Behavioral Signal Branch

The Behavioral Branch constructs a longitudinal account-level feature vector capturing the customer's return history patterns over their full account lifetime and over rolling time windows. Features include lifetime return rate, return rate by product category, average days-to-return, variability in days-to-return, count of distinct return reasons used historically, proportion of returns resulting in full refund versus store credit, frequency of address or payment method changes, and cross-channel return behavior (online purchase combined with in-store return).

The behavioral profile is designed to capture escalation trajectories. A legitimate customer's return rate tends to be stationary over time; a fraudster's rate tends to increase as they test the system's response thresholds. The framework computes first-order differences in rolling 30-day return rates to detect upward trajectories, and flags accounts where the behavioral profile has shifted substantially within the prior 90 days relative to the prior 12-month baseline.

Account age is treated as a moderating variable rather than a direct feature. A 40% return rate in an account's first two weeks is qualitatively different from the same rate after two years of purchase history. The Behavioral Branch conditions all aggregated features on account tenure quintile, enabling the meta-learner to learn tenure-specific risk thresholds from training data.

6.4. NLP Signal Branch

The NLP Branch analyzes the free-text return reason provided by customers at return initiation. Despite the 76% documented embellishment rate [4], no existing fraud detection system exploits this signal. Three categories of linguistic features are extracted.

First, reason-specificity features measure the information density of the return explanation. Genuinely defective merchandise tends to generate specific, mechanistic descriptions, while fraudulent wardrobing tends to produce vague formulations such as "changed my mind" or "didn't work for my needs." Specificity is operationalized through named entity density, noun phrase complexity, and the presence of product-specific technical vocabulary.

Second, consistency features cross-reference return reason text against the purchase category and product description. A return reason describing a sizing issue on an electronics purchase, or a defect claim on a consumable item, signals inconsistency that warrants elevated scrutiny. These features are computed using cosine similarity between the

return reason embedding and the product description embedding, both encoded with a pre-trained sentence transformer.

Third, pattern-repetition features detect templated or recycled language. Fraudsters who submit high volumes of fraudulent returns frequently reuse the same formulations or cycle through a small set of phrases. The NLP Branch computes Jaccard similarity between the current return reason and the customer's prior return reason history, flagging high-similarity returns as potentially formulaic.

The NLP Branch uses a fine-tuned BERT-base model as its core encoder. Pre-training on the masked language modeling objective provides strong general text representations; fine-tuning on a labeled return reason dataset adapts these representations to the specific vocabulary and fraud-relevant patterns of retail return text. The branch outputs both a fraud probability and a compressed embedding that the meta-learner uses directly as input features. The fine-tuning protocol described above is conditional on the availability of a labeled return reason corpus, which does not currently exist as a public resource. Section IV-B details a two-stage dataset strategy – beginning with a semi-synthetic benchmark and progressing to a retailer partnership – that is designed to produce the training data this branch requires. In the interim, the NLP branch can be initialized using a general-purpose sentence transformer without domain fine-tuning, with the understanding that its discriminative performance will be materially lower until a labeled corpus is available for adaptation.

6.5. Late-Fusion Stacking Architecture

The three branch outputs probability estimates, confidence scores, and for the NLP branch a compressed embedding – are passed to a gradient-boosted stacking meta-learner. The meta-learner is trained on a held-out validation set using branch outputs as features, following standard stacking protocols to prevent data leakage between branch training and meta-learner training.

Branch confidence scores modulate each branch's effective contribution. Low-confidence NLP outputs, for example, when a customer provides no free-text reason, reduce that branch's weight in the meta-learner's input space without requiring hard imputation. This is implemented by including a per-branch confidence flag as an explicit meta-learner input feature, allowing the meta-learner to learn appropriate down-weighting implicitly from training data.

The stacking architecture produces a final fraud probability and a severity-weighted risk score. The risk score is computed as the product of the fraud probability and the return value, approximating the expected financial exposure of the event. This score drives cost-sensitive threshold optimization as described in the following subsection.

Confidence scores are derived separately for each branch. For the Transaction and Behavioral branches, both gradient-boosted ensembles, confidence is computed as the standard deviation of predicted probabilities across individual trees: a tight distribution indicates high agreement among estimators and yields a high confidence score, while a dispersed distribution signals uncertainty and reduces the branch's effective contribution. For the NLP branch, confidence is derived from the entropy of the BERT-base output distribution over the binary fraud/non-fraud classes: low entropy indicates a confident prediction, while high entropy common when return reason text is absent or uninformative flags the output for down-weighting. In all three cases, confidence scores are normalized to the unit interval before being passed to the meta-learner as explicit input features.

6.6. Cost-Sensitive Learning and Threshold Optimization

Standard fraud classifiers minimize cross-entropy loss, which treats all misclassification errors equally. For return fraud, this is economically incorrect: a false negative on an $800 electronics return carries forty times the financial cost of a false negative on a $20 apparel return. Similarly, a false positive on a high-lifetime-value customer carries higher business cost than a false positive on a first-time buyer.

The framework adopts the example-dependent cost-sensitive framework of Bahnsen et al. [17][18], defining four cost terms for each transaction: the cost of a false negative (fraud value lost), the cost of a false positive (customer relationship cost, estimated as a fraction of customer lifetime value), and zero-cost terms for true positives and true negatives. The meta-learner is trained to minimize expected cost rather than cross-entropy, and the classification threshold is optimized to maximize savings, the net financial benefit of detection relative to a no-model baseline.

SMOTE [20] addresses class imbalance during branch-level training by generating synthetic minority-class examples in feature space rather than simply over-weighting observed fraud events. Ensemble SMOTE variants from Fernández et al. [21] are applied within each branch independently, as the appropriate sampling ratio may differ across feature modalities.

7. Proposed Evaluation Methodology

To validate the proposed framework, a comprehensive evaluation methodology must assess detection accuracy, system performance, and business impact across realistic operating conditions. The methodology is structured into five components.

7.1. Evaluation Metrics

The evaluation framework employs a layered metric structure reflecting both classification performance and financial impact. Standard detection metrics are defined using the formulas below:

$$Sensitivity = TP / (TP + FN)$$
$$Specificity = TN / (TN + FP)$$
$$PPV = TP / (TP + FP)$$
$$NPV = TN / (TN + FN)$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The framework targets sensitivity above 90% for high-severity fraud (return value exceeding one standard deviation above the mean) and specificity above 85% across all fraud categories.

AUC-PR is used as the primary scalar summary metric in preference to AUC-ROC, following Saito and Rehmsmeier [22], who demonstrate that precision-recall curves provide more informative comparisons under class imbalance. AUC-ROC is retained as a secondary metric for comparability with prior literature. The primary business metric is savings, defined as the sum of fraud values correctly identified minus the sum of customer lifetime value fractions lost to false positives. Customer lifetime value fraction is set at 5% as a default, with sensitivity analysis across the range of 1% to 15% to assess metric robustness to this assumption.

7.2. Dataset Strategy

No public return fraud dataset currently exists. The evaluation strategy addresses this gap through a two-stage approach. Stage 1 constructs a semi-synthetic benchmark by augmenting the IEEE-CIS Fraud Detection dataset with return-specific synthetic features generated from the empirical distributions reported in industry surveys [1][4][5]. Return reason text is generated using a template-based approach seeded with category-specific return reason taxonomies, with fraud examples overrepresenting vague formulations and legitimate examples overrepresenting specific product descriptions. This semi-synthetic benchmark enables initial validation of the framework's architecture and feature engineering choices.

Stage 2 pursues partnership with one or more large-volume retailers for access to anonymized transaction logs with labeled return outcomes. Required data elements include order-level purchase and return records spanning a minimum of 12 months and 1 million transactions, customer-level behavioral history, free-text return reason fields, and ground-truth fraud labels derived from merchandise inspection outcomes. This data enables full empirical validation under realistic distribution shifts and labeling noise.

7.3. Temporal Validation Protocol

Following the realistic fraud modeling methodology of Dal Pozzolo et al. [9], the evaluation employs strict temporal splitting rather than random train-test partitioning. The dataset is divided chronologically into training, validation, and test windows. The test window begins at least 90 days after the end of the training window to simulate production conditions and account for labeling delay, as fraud determinations based on merchandise inspection may not be finalized for weeks after the return event.

Cross-validation uses a sliding window protocol: multiple non-overlapping test windows are evaluated, each preceded by a training window of fixed duration, to assess performance stability across time periods. Models that perform well on randomly split data but degrade on temporal splits are identified and excluded from the final comparison, following the evaluation discipline Dal Pozzolo et al. established on 75 million real-world transactions over three years [9].

7.4. Concept Drift Evaluation

The four drift types of Gama et al. [19] sudden, gradual, incremental, and recurring are each tested explicitly. Sudden drift is simulated by introducing abrupt shifts in fraud prevalence or fraud type distribution at a known point in the test window. Gradual and incremental drift are simulated by progressively shifting feature distributions over the test window. Recurring drift tests the model's ability to recover performance after a policy change that temporarily reduces fraud prevalence before it returns to baseline levels.

For each drift type, the evaluation measures detection latency (time from drift onset to measurable performance degradation), the magnitude of performance drop under drift, and the recovery time after retraining with new data. These metrics directly inform retraining frequency recommendations for production deployments.

7.5. Ablation Study Design

A systematic ablation study evaluates the marginal contribution of each signal branch. Six model variants are evaluated: Transaction Branch only; Behavioral Branch only; NLP Branch only; Transaction plus Behavioral; Transaction plus NLP; and the full three-branch system. Performance across these variants, measured on the same temporal test split, quantifies the incremental value of each signal modality and the NLP branch in particular, given its novelty in this domain.

8. Discussion

8.1. Framework Advantages

The multi-signal framework addresses three specific limitations of existing fraud detection approaches as applied to the return fraud problem. First, the behavioral branch captures the longitudinal escalation patterns that characterize systematic return abusers, which single-event classifiers cannot detect by construction. A customer who returns 40% of purchases over 90 days does not exhibit any single anomalous event – only the account-level profile reveals the pattern.

Second, the NLP branch exploits a signal channel structurally unique to the return fraud context. In payment fraud, the fraudulent actor is typically absent from the transaction text there are no customer-supplied narratives to analyze. In return fraud, the customer actively constructs a justification, and the linguistic properties of that justification carry discriminative signal that the 76% embellishment rate [4] confirms is systematically distorted. No prior fraud detection framework, academic or commercial, has exploited this signal.

Third, the late-fusion stacking architecture provides operational modularity absent from end-to-end deep learning approaches. When fraud patterns shift in one modality for example, when fraudsters begin using more specific language in return reasons after learning that vague language triggers alerts only the NLP branch requires retraining. The meta-learner automatically adapts its branch weighting to the updated branch output without requiring full system retraining.

8.2. Limitations

Several limitations constrain the framework's applicability. The NLP branch depends on the availability and quality of free-text return reason data. Many retailers use dropdown-only return interfaces that generate categorical codes rather than natural language text; these customers' return events cannot benefit from NLP analysis. The proportion of returns with free-text reasons varies substantially across retail verticals, which limits the generalizability of NLP-dependent performance claims.

Ground truth labeling for return fraud is inherently noisy. Physical inspection of returned merchandise, the gold standard for fraud determination, covers only a subset of returns; the majority are relabeled and restocked without systematic inspection. Models trained on inspection-derived labels may learn to detect fraud that happens to be inspected rather than fraud that happens to occur, introducing selection bias into the training data. Empirical validation studies should document inspection rates and test whether model performance degrades when evaluated on unselected return samples.

The cost-sensitive learning framework requires calibrated estimates of customer lifetime value and customer relationship cost for false positives. These values are retailer-specific and may be difficult to obtain in academic collaboration settings. Sensitivity analysis across a range of assumed CLV values is a practical substitute for precise estimates but introduces additional uncertainty into the savings metric.

A fourth limitation concerns model interpretability. Retail fraud analysts require case-level explanations before acting on a flagged account particularly when the remediation involves restricting a customer's return privileges or escalating to manual review. The Transaction and Behavioral branches, both built on gradient-boosted ensembles, offer partial interpretability through feature importance rankings and SHAP values, which can identify which account-level signals for example, elevated return rate in the prior 30 days or anomalous days-to-return variance drove the branch prediction. The NLP branch presents a harder problem: BERT-based embeddings do not yield intuitive feature-level explanations, and attention weights, while extractable, have been shown to correlate inconsistently with model behavior. The meta-learner can attribute final risk scores across branches, which provides a coarse signal (e.g., "driven primarily by behavioral and NLP signals"), but it cannot expose the specific textual patterns that activated the NLP branch. Future work should evaluate attention visualization methods and integrated gradients as potential remedies for this gap, and consider whether a simpler, interpretable NLP alternative – such as logistic regression over explicit linguistic features – provides sufficient discriminative signal with materially better explainability.

8.3. Relationship to Existing Fraud Detection Infrastructure

Most large retailers operate third-party return fraud services such as Appriss Retail's Verify-1, which maintain cross-retailer fraud registries and apply rule-based thresholds. The proposed framework is designed to complement, not replace, these services. The behavioral and NLP branches surface signals that cross-retailer blacklists cannot detect account-level escalation patterns within a single retailer, and linguistic patterns in return narratives while the transaction branch provides a calibrated risk score that can be fused with the output of existing third-party systems. This positioning reduces the barrier to adoption by framing the framework as an additive layer over current infrastructure..

9. Conclusion and Future Work

This paper has characterized return fraud as a structurally distinct problem from payment fraud one defined by authenticated actors, extended temporal windows, and delayed ground truth labeling and demonstrated that these properties require dedicated architectural choices rather than direct adaptation of existing payment fraud methods. Building on this characterization, the paper introduced a multi-signal framework that merges transactional features, behavioral history profiles, and NLP analysis of return reason text through a late-fusion stacking architecture with cost-sensitive learning. The framework addresses a $103 billion annual problem that the fraud detection literature has systematically ignored and provides the first formal treatment of return reason text as a discriminative fraud signal.

Instead of relying on random-split comparisons common in the literature on payment fraud, this evaluation approach, based on temporal validation and example-dependent cost metrics, brings the methodology above, providing results that are more appropriate for deployment decisions in reality. The ablation study design specifically isolates the contribution of each signal branch, meaning that it's a strong test of the NLP branch value claim.

There are three aspects which future work should focus on. First, empirical verification with reference to actual retailer transactions with ground-truth fraud labels is needed to verify that the performance targets of the framework are feasible in real life and to calibrate the NLP branch on genuine return reason text at scale. Second, it would be beneficial to develop the framework to include cross-channel behavioral cues in-store purchase patterns, loyalty program data, customer service contact history that serve to enhance the contextualization of legitimate and fraudulent return behavior. Third, concept drift evaluation protocol should be

extended to include online learning methods that can dynamically adjust from production feedback, such that retraining lag time between the training of the model and the actual testing leads to performance gaps in case fraud behavior evolves quickly. As the ratio of return fraud to retail return burden grows, it is operationally critical to pursue structured machine learning approaches to return fraud. The multi-signal framework proposed here offers a principled starting point for this endeavor.

References

[1] Appriss Retail and Deloitte, "2024 Consumer Returns in the Retail Industry," Appriss Retail, Dec. 2024. https://apprissretail.com/resources/2024-consumer-returns-report/

[2] National Retail Federation, "NRF and Appriss Retail Report: $743 Billion in Merchandise Returned in 2023," NRF, Dec. 26, 2023. https://nrf.com/media-center/press-releases/nrf-and-appriss-retail-report-743-billion-merchandise-returned-2023.

[3] National Retail Federation and Happy Returns, "2025 Retail Returns Landscape," NRF, 2025. https://nrf.com/research/2025-retail-returns-landscape.

[4] Optoro, "With Returns Fraud & Abuse on the Rise, and 69% of Shoppers Admitting to Wardrobing, with 64% Doing So At Least Once a Month," Optoro, 2024. https://www.optoro.com/returns-news/optoro-launches-returns-unwrapped/.

[5] National Retail Federation, "NRF and Happy Returns Report: 2024 Retail Returns to Total $890 Billion

[6] ," NRF, Dec. 5, 2024. https://nrf.com/media-center/press-releases/nrf-and-happy-returns-report-2024-retail-returns-total-890-billion

[7] A. Mutemi and F. Bação, "E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review," Big Data Mining and Analytics, vol. 7, no. 2, pp. 419–444, Jun. 2024, doi: https://doi.org/10.26599/BDMA.2023.9020023.

[8] Y. Zhu et al., "Modeling Users' Behavior Sequences with Hierarchical Explainable Network for Cross-domain Fraud Detection," in Proc. The Web Conf. 2020 (WWW '20), Apr. 2020, pp. 928–938, doi: https://doi.org/10.1145/3366423.3380172.

[9] G. Zhang et al., "eFraudCom: An E-commerce Fraud Detection System via Competitive Graph Neural Networks," ACM Transactions on Information Systems, vol. 40, no. 3, pp. 1–29, Jul. 2022, doi: https://doi.org/10.1145/3474379.

[10] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784–3797, Aug. 2018, doi: https://doi.org/10.1109/tnnls.2017.2736643.

[11] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing Graph Neural Network-based Fraud Detectors against Camouflaged Fraudsters," Proc. 29th ACM Int. Conf. Inf. Knowl. Manag. (CIKM '20), Oct. 2020, doi: https://doi.org/10.1145/3340531.3411903.

[12] Y. Liu et al., "Pick and Choose: A GNN-based Imbalanced Learning Approach for Fraud Detection," Proceedings of the Web Conference 2021, Apr. 2021, doi: https://doi.org/10.1145/3442381.3449989.

[13] N. Tax et al., "Machine Learning for Fraud Detection in E-Commerce: A Research Agenda," Deployable Machine Learning for Security Defense, vol. 1482, pp. 30–54, 2021, doi: https://doi.org/10.1007/978-3-030-87839-9_2.

[14] J. Yang, K. Chen, K. Ding, C. Na, and M. Wang, "Auto Insurance Fraud Detection with Multimodal Learning," Data Intelligence, vol. 5, no. 2, pp. 388–412, Feb. 2023, doi: https://doi.org/10.1162/dint_a_00191.

[15] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: transformer based context aware online recruitment fraud detection," Discover Computing, vol. 28, no. 1, Feb. 2025, doi: https://doi.org/10.1007/s10791-025-09502-8.

[16] V. R. Saddi, B. Gnanapa, S. Boddu, and J. Logeshwaran, "The Role of Natural Language Processing in Detecting Insurance Fraud," 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216), pp. 1–6, Dec. 2023, doi: https://doi.org/10.1109/c2i659362.2023.10430658.

[17] Z. Zhang et al., "Identifying E-Commerce Fraud Through User Behavior Data: Observations and Insights," Data Science and Engineering, vol. 10, pp. 24–39, Jan. 2025, doi: https://doi.org/10.1007/s41019-024-00275-6.

[18] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134–142, Jun. 2016, doi: https://doi.org/10.1016/j.eswa.2015.12.030.

[19] A. C. Bahnsen, A. Stojanovic, D. Aouada and B. Ottersten, "Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk," 2013 12th International Conference on Machine Learning and Applications, Miami, FL, USA, 2013, pp. 333-338, doi: https://doi.org/10.1109/icmla.2013.68.

[20] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, Article 44, pp. 1–37, Mar. 2014, doi: https://doi.org/10.1145/2523813.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: https://doi.org/10.1613/jair.953.

[22] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, Apr. 2018, doi: https://doi.org/10.1613/jair.1.11192.

[23] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: https://doi.org/10.1371/journal.pone.0118432.