JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Beyond Binary Fraud Detection: Amount-Aware Operational Ranking for Transaction Risk Prioritization

Hartatik *

Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Sleman 55283, Yogyakarta, Indonesia; e-mail : hartatik@amikom.ac.id
* Corresponding Author : Hartatik

Abstract: Fraud detection in digital financial transactions is traditionally formulated as a binary classification problem, although real-world fraud investigation requires analysts to prioritize a limited number of suspicious transactions according to operational risk and potential financial impact. This study reformulates fraud detection as an amount-aware operational ranking problem for fraud-risk prioritization. Transactions are organized into time-window query groups, and fraudulent transactions are assigned graded relevance based on training-only transaction-amount quartiles, enabling the ranking objective to distinguish low- and high-severity fraud without relying on proprietary cost matrices. The proposed formulation is implemented using a representative Learning-to-Rank framework based on LambdaMART, while an out-of-fold XGBoost risk score is incorporated as an auxiliary feature to refine the ranking representation rather than serve as the primary contribution. Experiments conducted on a public credit-card fraud dataset using chronological validation and future-holdout testing demonstrate that amount-aware relevance consistently improves severity-aware top-rank ordering compared with conventional binary relevance. The proposed HybridLTR_amount model significantly outperforms XGBClassifier and PureLTR_binary in terms of all-query NDCG@10, whereas its performance is not statistically different from PureLTR_amount, indicating that the primary empirical improvement is attributable to the amount-aware ranking formulation rather than the auxiliary hybrid component. Additional operational analyses show that high-risk transactions and fraudulent financial losses are concentrated within a compact top-ranked segment, while budget-oriented evaluation demonstrates the practical value of the proposed formulation under limited analyst review capacity. These findings establish amount-aware operational ranking as an effective formulation-centric framework for operational fraud-risk prioritization rather than as a new classification algorithm.

Keywords: Amount-Aware Operational Ranking; Binary Classification; Financial Fraud Detection; Fraud-Risk Prioritization; Investigation Budget; Learning-to-Rank; Operational Risk Assessment; Transaction Fraud Detection.

Received: May, 24th 2026
Revised: July, 20th 2026
Accepted: July, 21st 2026
Published: July, 30th 2026

1. Introduction

Digital financial transactions have become a fundamental component of modern payment ecosystems. Their rapid growth has substantially increased transaction volume, processing speed, and service accessibility while simultaneously expanding opportunities for increasingly sophisticated fraudulent activities that must be detected under strict operational constraints [1]–[3]. Consequently, financial fraud detection has become a critical research area, with recent advances focusing on improving predictive performance through machine learning, deep learning, and hybrid intelligence models. Despite these developments, most existing studies continue to formulate transaction fraud detection as a binary classification problem, in which each transaction is independently classified as either fraudulent or legitimate [1], [2], [4]. Although this formulation provides reliable transaction-level risk estimation, it does not fully reflect the operational requirements of real-world fraud investigation.

The primary limitation of binary fraud classification is that it assumes all fraudulent transactions are equally important. Under this formulation, a low-value fraudulent transaction

receives the same positive label as a high-value fraudulent transaction, despite the substantial difference in their financial consequences. This creates a fundamental mismatch between the optimization objective of conventional fraud classifiers and the decision-making process followed by fraud analysts. In operational environments, investigators typically work under limited review capacity and must determine not only whether a transaction is fraudulent but also which suspicious transactions should be investigated first. Consequently, high classification performance, as measured by metrics such as AUC or AUPRC, does not necessarily imply that the most financially consequential fraud cases appear at the top of the analyst review list.

Beyond this objective mismatch, operational fraud detection is further complicated by severe class imbalance and temporal distribution shift. Fraudulent transactions usually constitute only a small proportion of all transactions, making accuracy-oriented evaluation potentially misleading and encouraging models to favor the dominant legitimate class [2], [4], [5]. Moreover, transaction behavior, customer activity, merchant characteristics, and fraud strategies evolve continuously over time, causing the statistical properties of future data to differ from historical observations [2], [6]. These characteristics suggest that fraud detection should not be viewed solely as a retrospective classification problem but also as a future-oriented prioritization problem that must remain effective under changing operational conditions.

These limitations motivate a shift from binary prediction toward operational fraud-risk prioritization. Rather than independently estimating the probability of fraud for each transaction, this study reformulates fraud detection as an amount-aware operational ranking problem. Transactions are grouped into time-based operational query windows, and fraudulent transactions are assigned graded relevance according to their financial severity. High-impact fraud therefore receives greater relevance than low-impact fraud, allowing the learning objective to prioritize financially consequential cases while remaining consistent with practical analyst review constraints. This formulation explicitly separates fraud identification from investigation priority, thereby aligning model optimization with the operational objective of minimizing financial loss under limited investigation capacity.

Learning-to-Rank (LTR) provides an appropriate learning paradigm for this formulation because it directly optimizes the relative ordering of candidate items within a query rather than predicting independent labels [7]. In the proposed framework, each operational time window is treated as a query, transactions within the window are treated as candidate items, and amount-aware fraud severity serves as graded relevance. LambdaMART is selected as the ranking model because it has demonstrated strong performance on structured tabular data and directly optimizes ranking quality through the rank:ndcg objective [7], [8]. It should be emphasized that this study does not introduce a new Learning-to-Rank algorithm. Instead, LambdaMART serves as a representative implementation for evaluating whether an amount-aware operational ranking formulation better supports fraud-risk prioritization than conventional binary classification or binary-relevance ranking.

To further strengthen the ranking representation, the proposed framework incorporates an out-of-fold XGBoost classifier score as an auxiliary risk signal. XGBoost remains one of the most effective supervised learning algorithms for structured tabular data because of its scalability, robustness, and predictive performance [9]. In this study, the classifier-derived score provides complementary evidence of transaction risk while reducing direct training leakage through chronological out-of-fold prediction. Nevertheless, the hybrid component is intentionally positioned as an auxiliary refinement rather than the primary scientific contribution. The central contribution lies in the reformulation of fraud detection as an amount-aware operational ranking problem and in demonstrating that the resulting formulation better supports operational fraud prioritization under temporal and investigation-budget constraints.

Based on this motivation, this study investigates the following research hypothesis: reformulating fraud detection as an amount-aware operational ranking problem provides more effective operational fraud-risk prioritization than conventional binary classification or binary-relevance ranking while remaining robust under realistic temporal deployment conditions. To evaluate this hypothesis, the proposed framework is assessed using operational ranking metrics, including NDCG@K, Recall@K, FraudAmountRecall@K, captured fraud amount, risk concentration, and investigation-budget analysis. Conventional classification metrics, including AUC and AUPRC, are retained only as supporting references for the classifier baseline. The experimental protocol further incorporates chronological data partitioning, future-holdout evaluation, paired block-bootstrap significance testing, query-window

sensitivity analysis, rolling-origin validation, temporal drift analysis, and ablation studies to comprehensively validate both the proposed formulation and its individual components.

The main contributions of this study are summarized as follows. First, this study reformulates transaction fraud detection from binary classification into an amount-aware operational ranking problem that explicitly aligns the learning objective with practical fraud investigation. Second, it introduces an amount-aware graded-relevance strategy that represents financial severity and overcomes the limitation of conventional binary fraud labels. Third, it implements the proposed formulation using a LambdaMART-based Learning-to-Rank framework enhanced by an out-of-fold XGBoost risk score as an auxiliary feature. Fourth, it evaluates the proposed formulation using operationally relevant ranking metrics and investigation-budget-oriented analyses rather than relying primarily on conventional classification performance. Finally, it demonstrates the robustness of the proposed formulation through temporal validation, risk-decile analysis, and ablation experiments, thereby distinguishing the contribution of amount-aware ranking from that of the auxiliary hybrid component.

2. Literature Review

2.1. Conventional Fraud Detection and Its Operational Limitations

Financial fraud detection has been extensively studied as a supervised learning problem in which each transaction is classified as either fraudulent or legitimate. Early research primarily relied on rule-based systems and statistical models, whereas recent advances have increasingly adopted machine learning, ensemble learning, deep learning, graph-based learning, self-supervised learning, and federated learning to improve fraud detection performance [1]–[3], [5], [10]–[12]. Collectively, these developments have significantly enhanced the capability of fraud detection systems to model complex transaction patterns, mitigate class imbalance, capture temporal behavior, and exploit relational dependencies among financial entities.

Despite these methodological advances, most existing studies continue to formulate fraud detection as a binary classification problem. Under this formulation, the model predicts whether an individual transaction belongs to the fraud class, and its performance is typically evaluated using classification-oriented metrics such as accuracy, precision, recall, F1-score, AUC, and AUPRC [1], [2], [4]. Although this paradigm is effective for transaction-level discrimination, it does not fully represent the operational workflow of fraud investigation. In practice, fraud analysts must determine not only whether a transaction is suspicious but also which suspicious transactions should be investigated first when review resources are limited.

This formulation introduces two important operational limitations. First, all fraudulent transactions receive the same positive label regardless of their financial impact. Consequently, low-value and high-value fraud cases are treated equally, even though their operational consequences may differ substantially. This creates a severity gap between binary fraud labels and the actual financial risk faced by investigators. Second, binary classifiers optimize transaction-level discrimination rather than investigation-oriented prioritization. As a result, a model may achieve high AUC or AUPRC while still failing to rank the most financially consequential fraud cases near the top of the analyst review queue. Therefore, strong classification performance does not necessarily translate into effective operational fraud-risk prioritization.

Another limitation arises from the dynamic nature of financial fraud. Real-world fraud detection systems operate under continuously evolving transaction behavior, customer activity, merchant characteristics, seasonal patterns, and adversarial fraud strategies [4], [6]. Recent surveys have consequently highlighted that modern fraud detection must address not only severe class imbalance but also temporal drift, dataset limitations, interpretability, and realistic evaluation protocols [4], [13]. These observations indicate that conventional random data splitting and classification-oriented metrics are insufficient for assessing deployment performance. Instead, chronological validation, future-holdout testing, temporal drift analysis, and budget-aware evaluation are required to determine whether a fraud detection model remains effective under future operational conditions.

Taken together, the current literature suggests that the principal challenge is no longer the development of another high-performing fraud classifier. Rather, the fundamental issue lies in the mismatch between the optimization objective of binary classification and the operational requirement to prioritize suspicious transactions under limited investigation capacity. This observation motivates the reformulation proposed in this study, in which fraud detection

is viewed as an amount-aware operational ranking problem rather than solely as a binary classification task.

2.2. Cost-Sensitive, Severity-Aware, and Operational Fraud Decision-Making

Recognizing the limitations of conventional binary classification, several studies have sought to make fraud detection more operationally meaningful by incorporating class imbalance, misclassification costs, financial losses, and decision consequences into the learning process [3], [4], [6], [13]. Among these approaches, cost-sensitive learning has become one of the most widely adopted strategies because false positives and false negatives impose substantially different operational costs. Missing a high-value fraudulent transaction may result in considerable financial loss, whereas reviewing excessive false positives increases analyst workload and operational expenses. By assigning different penalties to different prediction errors, cost-sensitive learning extends the conventional classification framework toward more practical decision-making.

Nevertheless, cost-sensitive classification remains fundamentally a transaction-level prediction problem. Although it modifies the learning objective or decision threshold, it does not explicitly optimize the order in which suspicious transactions should be reviewed. In operational fraud investigation, analysts typically work with a constrained review budget, making prioritization as important as classification accuracy. Under these circumstances, the practical objective is to maximize the financial value of detected fraud within limited investigative capacity rather than simply minimize classification errors. Consequently, a statistically accurate classifier may still be operationally inefficient if high-impact fraud cases are not concentrated within the top-ranked portion of the review queue.

Severity-aware and amount-aware fraud modeling further extend this perspective by recognizing that fraudulent transactions differ considerably in their financial consequences. Instead of treating every fraud instance as an equally important positive sample, these approaches incorporate transaction amount or estimated financial loss into either the learning objective or the evaluation process [4], [6], [14]. Such considerations are closely aligned with operational fraud investigation, where transactions associated with greater financial risk should receive higher investigation priority. However, most existing studies continue to evaluate performance primarily through classification-oriented metrics, while financial severity and investigation-budget constraints remain secondary considerations.

These observations reveal an important research opportunity. Existing studies acknowledge that fraud severity influences operational decision-making, yet most approaches continue to formulate the problem as weighted binary classification rather than explicit prioritization. This study builds upon the severity-aware perspective but advances it by reformulating fraud detection as a Learning-to-Rank problem. Rather than adjusting classification penalties, the proposed framework directly learns the investigation order of suspicious transactions according to their operational importance. This formulation is also consistent with recent developments in financial risk ranking, where ranking models are increasingly adopted to allocate limited attention and intervention resources to the highest-risk entities [14]. Based on this synthesis, we hypothesize that explicitly optimizing transaction ranking according to fraud severity provides a more operationally appropriate framework for fraud investigation than conventional cost-sensitive or binary classification approaches.

2.3. Learning to Rank for Financial Risk Prioritization

Learning to Rank (LTR) provides a natural learning paradigm for operational fraud-risk prioritization because it optimizes the relative ordering of candidate items within a query rather than predicting each transaction independently [7]. Originally developed for information retrieval, LTR learns to rank documents according to their relevance to a user query. The same principle can be extended to fraud investigation by treating each operational time window as a query and the transactions within that window as candidate items. Under this formulation, the learning objective shifts from transaction-level classification toward prioritizing transactions according to their investigation priority.

This ranking formulation is inherently more consistent with operational fraud investigation than conventional pointwise binary classification. Fraud analysts typically operate under limited review capacity, making the quality of the highest-ranked transactions more important than overall classification performance. Consequently, ranking-oriented metrics, including

NDCG@K, Recall@K, FraudAmountRecall@K, captured fraud amount, and risk concentration, provide a more meaningful assessment because they directly measure whether financially important fraud cases are concentrated within the available investigation budget [7], [8].

Recent studies in financial risk management further support this perspective by demonstrating that ranking-based models are preferable when limited investigative or intervention resources must be allocated to the highest-risk entities [14]. Although these studies address different financial applications, they share the same operational objective of prioritizing limited resources according to risk. This convergence suggests that fraud investigation can also benefit from a ranking-oriented formulation rather than relying solely on transaction-level classification.

Based on this rationale, the present study adopts LambdaMART as a representative Learning-to-Rank implementation because of its effectiveness on structured tabular data and its direct optimization of ranking quality through the rank:ndcg objective [7], [8]. Importantly, the contribution of this study does not lie in proposing a new ranking algorithm. Instead, LambdaMART serves as an implementation framework for evaluating the hypothesis that an amount-aware ranking formulation better supports operational fraud-risk prioritization than binary classification or binary-relevance ranking.

2.4. Temporal Robustness and Ablation-Oriented Evaluation

Operational fraud detection systems must remain effective under continuously evolving transaction environments rather than only on historical datasets. In practice, customer behavior, merchant activity, fraud prevalence, and adversarial strategies change over time, causing future transaction distributions to differ from those observed during model development [4], [6], [13]. Consequently, evaluation protocols based solely on static or randomly partitioned datasets may overestimate real-world deployment performance and provide an incomplete assessment of operational utility.

To better reflect practical deployment scenarios, recent studies increasingly advocate chronological validation, future-holdout testing, rolling-origin validation, and temporal drift analysis as more realistic evaluation strategies [4], [6], [13]. Unlike conventional random train-test splitting, these protocols assess whether a model maintains its effectiveness when confronted with future transaction streams and evolving fraud patterns. Such evaluations provide stronger evidence of model robustness and are therefore more representative of operational fraud detection systems.

Beyond robustness evaluation, it is equally important to determine which methodological component is responsible for any observed performance improvement. This issue is particularly relevant for hybrid frameworks, where performance gains may originate from multiple factors, including the learning formulation, relevance definition, feature representation, or auxiliary prediction signals. Recent studies have consequently emphasized the importance of ablation analysis to isolate the contribution of individual components rather than attributing improvements solely to the overall framework [15].

Collectively, these studies indicate that a convincing evaluation of operational fraud prioritization should assess not only predictive performance under temporal distribution shifts but also the contribution of each methodological component. Therefore, a comprehensive evaluation framework should combine deployment-oriented validation with component-wise analysis to provide reliable evidence of both model robustness and scientific contribution.

2.5. Research Gap and Positioning of This Study

The preceding review highlights substantial progress in financial fraud detection through advances in classification algorithms, representation learning, graph-based modeling, privacy-preserving learning, and robustness-oriented evaluation. As summarized in Table 1, these research directions have significantly improved transaction-level fraud prediction and addressed important challenges such as class imbalance, temporal distribution shifts, and model reliability. Nevertheless, their primary objective remains the accurate identification of fraudulent transactions rather than the operational prioritization of investigation resources.

Although recent studies have recognized the importance of financial severity and risk-aware decision-making, comparatively limited attention has been devoted to reformulating transaction fraud detection as a ranking problem that explicitly reflects investigation capacity and financial impact. Consequently, a model with excellent classification performance may

still provide limited operational value if financially significant fraud cases are not concentrated within the relatively small set of transactions that analysts can realistically review. This observation reveals a fundamental gap between conventional fraud prediction and practical fraud investigation.

Table 1. Positioning of the proposed study relative to major fraud-detection research directions.

Research direction	Primary objective	Typical formulation	Operational limitation	Position of this study
Rule-based and statistical methods [1]–[3]	Detect suspicious transactions	Rule-based or binary classification	Limited adaptability and no severity-aware prioritization	Uses data-driven ranking within operational time windows
Conventional machine learning [2], [9]	Improves transaction-level prediction	Binary classification	High classification performance does not ensure effective investigation prioritization.	Reformulates fraud detection as an operational ranking problem
Imbalance-aware learning [2], [4], [5], [13]	Improve minority fraud detection	Binary classification or representation learning	Treats all fraud cases equally	Introduces amount-aware graded relevance
Deep and sequential learning [9], [13], [16], [17]	Capture nonlinear and temporal patterns	Classification or anomaly detection	Does not explicitly optimize investigation order	Learns ranking within operational query windows
Graph-based and explainable methods [10], [11], [17]–[25]	Exploit relational information	Graph-based classification	Focuses primarily on fraud detection rather than prioritization	Prioritizes investigation order instead of relational modeling
Federated and privacy-aware learning [26]	Privacy-preserving fraud detection	Distributed classification	Addresses deployment constraints but not investigation prioritization	Targets limited analyst review capacity
Distribution-shift and trustworthy learning [6]	Improve robustness	Classification with robustness evaluation	Limited emphasis on investigation utility	Incorporates chronological validation and temporal robustness analysis
Financial risk ranking [14]	Prioritize high-risk entities	Learning to Rank	Not specifically designed for transaction-level fraud prioritization	Extends ranking principles to amount-aware transaction fraud prioritization
Ablation-oriented evaluation [15]	Identify component contribution	Component-wise comparison	Rarely applied to fraud prioritization	Separates the contribution of amount-aware relevance from the auxiliary hybrid signal
Proposed framework	Prioritize high-impact fraud under limited review capacity	Time-window-based Learning to Rank with amount-aware graded relevance	Focuses on problem reformulation rather than a new ranking algorithm	Provides an operational formulation aligned with real investigation workflows

Motivated by this gap, this study hypothesizes that reformulating transaction fraud detection as an amount-aware operational ranking problem provides a learning objective that is more consistent with operational fraud investigation than conventional binary classification or binary-relevance ranking. Rather than treating all fraudulent transactions as equally important, the proposed formulation prioritizes suspicious transactions according to their financial severity while preserving realistic operational constraints.

Figure 1 presents the overall framework of the proposed amount-aware operational ranking formulation. The framework integrates leakage-safe feature construction, time-based query formation, amount-aware graded relevance, an auxiliary out-of-fold XGBoost risk signal, and a LambdaMART ranking model to generate prioritized transaction lists for operational fraud investigation. Importantly, the proposed contribution lies in the problem formulation and its operational evaluation rather than in the development of a new Learning-to-Rank algorithm.

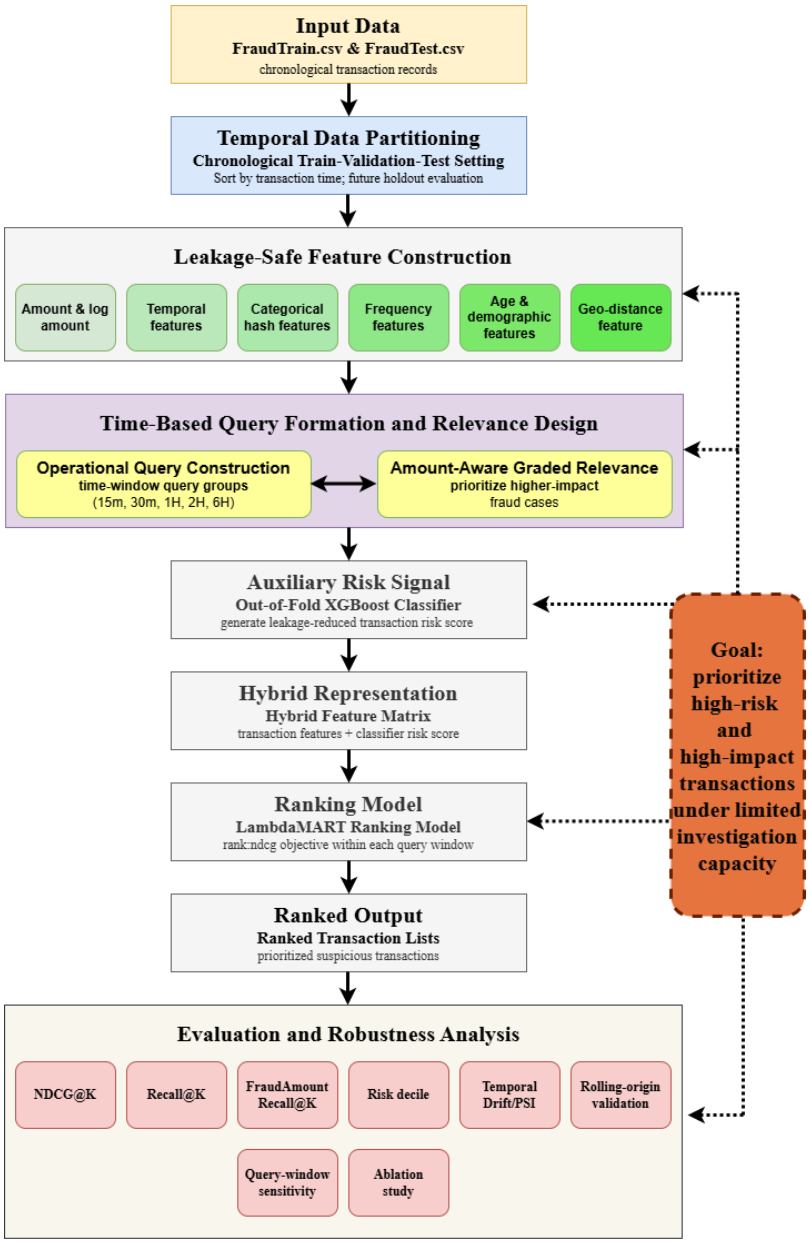

Figure 1. Amount-aware operational ranking formulation for transaction fraud-risk prioritization.

The proposed framework is subsequently evaluated using ranking-oriented performance measures, investigation-budget analysis, temporal robustness assessment, and component-wise ablation experiments to determine whether the proposed formulation provides meaningful operational advantages under realistic fraud investigation scenarios.

3. Proposed Method

This study implements transaction fraud detection as an amount-aware operational ranking framework rather than as a standalone binary classification task. Instead of optimizing transaction-level fraud prediction, the proposed framework prioritizes suspicious transactions

within time-based investigation windows according to both fraud risk and financial severity. This design enables the learning objective to better reflect practical fraud investigation, where analyst capacity is inherently limited and only a subset of transactions can be reviewed. Figure 1 presents the overall workflow of the proposed framework and illustrates how leakage-safe feature construction, query formation, relevance design, and ranking are integrated into a unified operational fraud-prioritization pipeline.

3.1. Implementation Scope and Design Rationale

The proposed framework is designed to evaluate a single research question: Can fraud detection better support operational investigation when transactions are ranked according to both fraud risk and financial severity within time-based review windows? This perspective differs from conventional fraud detection, which predicts each transaction independently and typically assigns identical importance to all fraudulent cases.

To address this question, the framework is constructed around three complementary design decisions. First, transactions are organized into time-based query windows, allowing the ranking model to learn the relative investigation priority of transactions within each operational review period instead of producing independent fraud scores. This query-based formulation reflects the way fraud analysts process transaction queues in practice.

Second, fraudulent transactions are represented using amount-aware graded relevance rather than binary labels. The graded relevance design distinguishes fraud cases according to their financial severity, enabling the ranking objective to prioritize transactions with greater operational impact while preserving the distinction between legitimate and fraudulent transactions.

Third, the experimental design explicitly isolates the contribution of each major component. Rather than evaluating only the complete framework, multiple model variants are introduced to determine whether observed improvements originate from the ranking formulation, the proposed relevance design, or the auxiliary classifier-derived risk signal. This formulation-oriented evaluation is intended to separate the contribution of the proposed learning objective from that of supporting implementation components.

Table 2 summarizes the experimental variants and their respective roles in validating the proposed framework. Together, these comparisons provide a structured evaluation of the proposed formulation from the perspectives of ranking strategy, relevance design, and auxiliary information integration.

Table 2. Experimental model variants and their roles in the evaluation.

Model variant	Purpose	Role in the evaluation
XGBClassifier	Conventional pointwise fraud classifier	Baseline for comparing binary classification with ranking-based prioritization
PureLTR_binary	LambdaMART using binary fraud relevance	Evaluates the effect of the ranking formulation without severity awareness
PureLTR_amount	LambdaMART using amount-aware graded relevance	Evaluates the proposed relevance formulation independently of the auxiliary classifier
HybridLTR_amount	PureLTR_amount enhanced with an out-of-fold XGBoost risk score	Assesses whether classifier-derived evidence further improves ranking performance
ShuffledHybridControl	Hybrid model with randomly shuffled auxiliary scores	Verifies that the auxiliary signal contributes meaningful information rather than random variation
AmountScore	Ranking based solely on transaction amount.	Determines whether performance can be explained by transaction amount alone
Random	Random transaction ordering	Lower-bound reference for ranking performance

3.2. Data Partitioning and Leakage-Safe Feature Construction

The proposed framework adopts a chronological data-partitioning strategy to approximate future deployment conditions and minimize information leakage between historical and

future transactions. The original training data are first sorted according to transaction time before being divided into training and validation subsets, while the provided test data are preserved as an independent future holdout set. The training subset is used for model fitting, whereas the validation subset supports model selection, hyperparameter tuning, and early stopping. The future holdout set is reserved exclusively for final performance evaluation. This temporal partitioning reflects realistic deployment scenarios in which future transactions are unavailable during model development and therefore avoids the optimistic bias that may arise from random data splitting.

Each transaction is represented by an engineered feature vector composed of six complementary feature groups. Rather than introducing new feature representations, the framework combines commonly used transactional, behavioral, spatial, and categorical descriptors to provide a comprehensive representation of transaction characteristics while maintaining leakage-safe feature construction. Table 3 summarizes the implemented feature groups and their respective purposes.

Table 3. Implemented feature groups.

Feature group	Implemented features	Purpose
Amount features	Raw amount and log-transformed amount	Capture transaction magnitude while reducing skewness
Temporal features	Hour, day of week, month, weekend indicator, and night indicator	Capture time-dependent behavioral patterns
Demographic features	Age and available customer attributes	Represent customer behavioral characteristics
Geographic features	Customer location, merchant location, and customer–merchant distance	Capture spatial inconsistency between customer and merchant
Categorical hash features	Merchant, category, gender, city, state, and job	Represent high-cardinality categorical variables without vocabulary refitting.
Frequency features	Training-only frequency encoding for selected categorical variables	Capture repeated behavioral patterns while preventing future information leakage

To maintain temporal consistency, all frequency encodings are estimated exclusively from the training subset before being applied to the validation and future holdout data. Similarly, stable hash encoding is adopted for high-cardinality categorical attributes so that previously unseen categories can be represented without reconstructing the encoding space. Other feature transformations, including logarithmic transformation of transaction amount and Haversine-based customer–merchant distance computation, are incorporated as standard implementation procedures and are therefore presented without further theoretical elaboration. When the auxiliary classifier-derived risk signal is incorporated, the resulting transaction representation is denoted by $\left(\vec{z_i}\right)$, which extends the engineered feature vector with the classifier-derived score. The construction of this hybrid representation and its integration into the proposed ranking framework are described in the following subsection.

3.3. Operational Query Construction and Amount-Aware Relevance Labeling

The proposed ranking task is constructed by grouping transactions into fixed time-based operational query windows, where all transactions within the same window are ranked against one another. This query-based formulation reflects the practical workflow of fraud investigation, in which analysts review batches of transactions generated during a specific operational interval rather than evaluating transactions independently. The primary experiments use a one-hour query window, as this interval provides a practical balance between transaction volume and analyst review capacity.

To evaluate whether the proposed formulation is robust to different operational settings, additional query-window sizes are considered during sensitivity analysis. Specifically, 15-minute, 30-minute, 1-hour, 2-hour, and 6-hour windows are evaluated to represent increasingly

restrictive and relaxed investigation intervals. i, the query identifier q_i is determined from the transaction timestamp t_i, the earliest timestamp $t_{\min}$, and the selected window size Δ:

$$q_i = \left\lfloor \frac{t_i - t_{\min}}{\Delta} \right\rfloor \tag{1}$$

Fraud severity is then represented through an amount-aware graded relevance scheme. Legitimate transactions are assigned relevance level 0, whereas fraudulent transactions receive graded relevance values from 1 to 4 according to transaction-amount quartiles estimated exclusively from fraudulent transactions in the training subset. Using training-only quartiles prevents information leakage while ensuring that the relevance definition remains consistent across validation and future test data. The relevance label is defined as:

$$r_i = \begin{cases} 0, & y_i = 0 \\ 1, & y_i = 1 \text{ and } a_i \le Q_{25} \\ 2, & y_i = 1 \text{ and } Q_{25} < a_i \le Q_{50} \\ 3, & y_i = 1 \text{ and } Q_{50} < a_i \le Q_{75} \\ 4, & y_i = 1 \text{ and } a_i > Q_{75} \end{cases} \tag{2}$$

where y_i denotes the fraud label, a_i represents the transaction amount, and Q_{25}, Q_{50}, and Q_{75} denote the fraud-amount quartiles computed exclusively from the training subset. This relevance design separates fraud identification from investigation priority by allowing transaction amount to determine the ranking priority only after fraud status has been established. Consequently, the ranking objective emphasizes financially significant fraud without modifying the underlying fraud labels.

To further enrich the transaction representation, the proposed framework incorporates an auxiliary classifier-derived risk score generated by an XGBoost classifier. Rather than serving as the final prediction, this score is used as an additional feature that complements the engineered transaction representation while preserving the amount-aware ranking objective. To prevent information leakage, auxiliary scores for the training subset are generated using chronological out-of-fold prediction based on three time-series folds. In each fold, the classifier is trained on earlier observations and predicts the immediately following chronological segment. For the validation and future holdout sets, a final classifier trained on the complete training subset is used to generate the corresponding risk scores. This procedure ensures that every auxiliary prediction is obtained without access to future observations. The hybrid transaction representation is therefore defined as:

$$\vec{z}_i = [\vec{x}_i, \hat{p}_i] \tag{3}$$

where $\vec{x}_i$ denotes the engineered transaction feature vector and $\hat{p}_i$ denotes the auxiliary XGBoost-derived fraud-risk score. The resulting representation is subsequently optimized using a LambdaMART ranker, where each query corresponds to a time window, each item corresponds to a transaction, and the graded relevance label represents fraud severity. The implementation details of the auxiliary classifier and ranking model are summarized in Table 4.

To distinguish the contribution of each component, several ranking variants are implemented using the same Learning-to-Rank framework. PureLTR_binary evaluates the ranking formulation with binary relevance, PureLTR_amount evaluates the proposed amount-aware relevance independently of the auxiliary classifier, HybridLTR_amount assesses the additional value of classifier-derived risk information, and ShuffledHybridControl replaces the auxiliary score with a randomly shuffled counterpart to verify that any improvement originates from meaningful predictive information rather than random variation. The principal implementation parameters used throughout these experiments are summarized in Table 5.

The scale_pos_weight parameter is computed automatically from the class distribution of the training subset and is applied only to the auxiliary classifier. In contrast, the LambdaMART model is optimized directly using query groups and amount-aware relevance labels rather than classifier probabilities. After training, each transaction receives a ranking score within its corresponding query window, and transactions are ordered in descending score to produce the final investigation priority list. This design ensures that the classifier-derived score serves only as complementary evidence, while the final prioritization remains governed by the proposed amount-aware ranking formulation.

Table 4. Implementation of the auxiliary classifier and LambdaMART ranking framework.

Component	Implementation
Auxiliary classifier	XGBoost binary classifier
Auxiliary objective	binary: logistic
Training score generation	Chronological out-of-fold prediction
Number of out-of-fold splits	3
Validation and test score generation	Final classifier trained on the complete training subset
Ranker input	Engineered transaction features + auxiliary classifier-derived score
Ranking model	LambdaMART implemented using the XGBoost rank:ndcg objective.
Query definition	Time-based operational query window
Ranking item	Individual transaction
Relevance label	Amount-aware graded relevance $(r_i \in 0,1,2,3,4)$
Training evaluation metric	NDCG@10, NDCG@50
Model output	Ranked transaction list within each query window

Table 5. Main configuration of the XGBoost classifier and LambdaMART ranker.

Component	Parameter	Value
Auxiliary XGBoost	Objective	binary: logistic
	Evaluation metrics	AUC, AUPRC
	Boosting rounds	450
	Early stopping	40
	Learning rate	0.035
	Maximum depth	5
	Minimum child weight	5
	Subsample/column subsample	0.85 / 0.85
	Regularization	$(\lambda = 2.0, \alpha = 0.2)$
	Tree method	hist
	Class imbalance	scale_pos_weight
Auxiliary score	OOF strategy	Time-series split (3 folds)
LambdaMART	Objective	rank:ndcg
	Evaluation metrics	NDCG@10, NDCG@50
	Boosting rounds	450
	Early stopping	40
	Learning rate	0.045
	Maximum depth	5
	Minimum child weight	10
	Subsample/column subsample	0.85 / 0.85
	Regularization	$(\lambda = 2.0, \alpha = 0.1)$
	Tree method	hist
Random seed	Base seed	42

3.4. Evaluation Design

The evaluation framework is designed to determine whether the proposed ranking formulation improves operational fraud prioritization under realistic analyst-capacity constraints. Because the objective of this study is to prioritize transactions within time-based investigation windows rather than to classify transactions independently, the evaluation primarily relies on ranking-oriented and budget-oriented measures. Conventional classification metrics, including AUC and AUPRC, are retained only as supporting references for the XGBoost classification baseline and are not used to assess the proposed ranking framework.

All experiments adopt a single all-query evaluation setting in which every operational query window is included, regardless of whether fraudulent transactions are present. This evaluation protocol reflects continuous fraud-monitoring environments, where a deployed system must rank transactions for every review interval without prior knowledge of fraud occurrence. Consequently, the evaluation measures both ranking effectiveness and operational consistency across the complete transaction stream.

To represent different levels of analyst capacity, investigation budgets of K = 5, 10, 20, 50, and 100 transactions per query are evaluated. Smaller budgets represent highly constrained investigation resources, whereas larger budgets approximate more relaxed review scenarios. The operational evaluation metrics used throughout the experiments are summarized in Table 6. Together, these metrics quantify ranking quality, fraud retrieval capability, investigation efficiency, and the financial value of fraud captured within limited review budgets.

Table 6. Operational evaluation metrics.

Metric	Evaluation objective	Operational interpretation
NDCG@K	Measures ranking quality using amount-aware graded relevance	Determines whether financially severe fraud cases appear near the top of the investigation list
Recall@K	Measures the proportion of fraudulent transactions retrieved within the review budget	Evaluates fraud coverage under limited analyst capacity
FraudAmountRecall@K	Measures the proportion of fraudulent transaction amount recovered within the review budget	Evaluates the ability to prioritize financially significant fraud
Precision@K	Measures the proportion of reviewed transactions that are fraudulent	Evaluates investigation efficiency
CapturedFraudAmount@K	Measures the average fraudulent transaction amount captured within the top-K transactions of each query	Quantifies the financial value recovered during operational investigation

Among these measures, NDCG@K serves as the primary evaluation metric because it directly reflects the proposed amount-aware graded relevance formulation. The remaining metrics provide complementary perspectives by measuring fraud coverage, review efficiency, and financial impact under different investigation budgets, thereby offering a more comprehensive assessment of operational performance than ranking quality alone.

To evaluate statistical significance, paired block-bootstrap resampling is performed over operational query windows. Each bootstrap block contains 24 consecutive query windows, corresponding approximately to one day under the primary one-hour query configuration, and the procedure is repeated 400 times. Performing resampling at the query level preserves the temporal dependency of operational transaction streams while providing paired comparisons between the proposed framework and the principal baseline and control variants.

3.5. Robustness, Drift, and Ablation Analysis

Beyond overall ranking performance, the proposed framework is further evaluated through a series of robustness and diagnostic analyses designed to examine its reliability under different deployment conditions and to identify the contribution of individual framework components. These complementary analyses assess score concentration, temporal stability, operational sensitivity, future generalization, and formulation-level contribution, thereby providing a comprehensive validation of the proposed ranking framework.

Risk-decile analysis evaluates whether fraudulent transactions are concentrated within the highest-ranked portion of the transaction stream. Test transactions are sorted according to their HybridLTR_amount ranking scores and divided into ten approximately equal-sized deciles, where the first decile represents the highest investigation priority. This analysis examines whether the proposed ranking formulation successfully concentrates fraudulent activity within the earliest review segments.

Temporal drift analysis investigates ranking stability under evolving transaction distributions. The future holdout period is partitioned into monthly intervals, and ranking

performance together with fraud prevalence is evaluated for each period. In addition, the Population Stability Index (PSI) is computed using ten bins to quantify distributional changes between the training and future holdout data. This analysis determines whether the proposed framework maintains consistent performance as transaction characteristics evolve over time.

Query-window sensitivity analysis examines whether the proposed formulation remains effective under different operational review intervals. Besides the primary one-hour configuration, additional window sizes of 15 minutes, 30 minutes, 2 hours, and 6 hours are evaluated to represent institutions with different transaction volumes, alert frequencies, and analyst workloads.

Rolling-origin validation provides an additional assessment of temporal robustness beyond a single chronological partition. Three expanding chronological folds are constructed so that each model is trained on earlier transactions, validated on the subsequent period, and evaluated on later observations. This future-oriented evaluation examines whether the proposed formulation consistently generalizes across multiple deployment periods.

Finally, ablation analysis is conducted to isolate the contribution of the principal components of the proposed framework. Rather than evaluating only the complete model, individual components are removed or modified while maintaining the same Learning-to-Rank framework. This experimental design enables the contribution of the amount-aware relevance formulation, the auxiliary classifier-derived score, and individual feature groups to be assessed independently. The evaluated ablation variants are summarized in Table 7.

Table 7. Ablation variants.

Ablation variant	Modification	Evaluation objective
ProposedFull	Complete framework with amount-aware relevance, auxiliary classifier score, and all engineered features	Reference implementation
NoHybridScore	Removes the auxiliary XGBoost-derived risk score	Evaluates the contribution of the auxiliary classifier
BinaryRelevance	Replaces graded relevance with binary fraud relevance	Evaluates the contribution of amount-aware relevance
NoHighCardinalityFrequency	Removes frequency-encoding features	Evaluates the importance of behavioral frequency information
NoTemporalFeatures	Removes temporal features	Evaluates the contribution of temporal transaction behavior
NoGeoDistanceFeatures	Removes geographic and distance-related features	Evaluates the contribution of spatial transaction information
NoAmountFeatures	Removes transaction-amount features while retaining graded relevance labels	Distinguishes the contribution of feature representation from relevance formulation
NoCategoricalHashFeatures	Removes hash-encoded categorical features	Evaluates the contribution of categorical transaction context
ShuffledHybridScore	Replaces the auxiliary classifier score with randomly shuffled values	Negative control for validating the auxiliary risk signal

For all trainable variants, the same Learning-to-Rank framework is retained to ensure fair comparison across ablation settings. Existing trained models are reused whenever an equivalent configuration already exists, while new models are trained using identical optimization settings. Each variant is evaluated using the same query-level metrics employed in the main experiments, including NDCG@K, Recall@K, FraudAmountRecall@K, and Precision@K, thereby ensuring that performance differences can be attributed directly to the modified component rather than to changes in the evaluation protocol.

4. Results and Discussion

4.1. Amount-Aware Operational Ranking as the Main Contribution

The primary contribution of this study is the reformulation of transaction fraud detection as an amount-aware operational ranking problem rather than the development of a new hybrid classification algorithm. This distinction reflects the practical nature of fraud

investigation, where analysts must prioritize suspicious transactions under limited review capacity instead of merely determining whether individual transactions are fraudulent. Accordingly, the proposed framework organizes transactions into time-based query windows and ranks them according to both fraud relevance and financial severity, thereby shifting the learning objective from independent pointwise classification toward investigation-oriented prioritization.

To evaluate this formulation under realistic deployment conditions, the experiments adopt a strictly chronological protocol. The original training file is divided into training and validation subsets, whereas the provided test file is retained as an independent future holdout period. In addition, the amount thresholds used to construct the graded relevance labels are estimated exclusively from fraudulent transactions in the training subset. This design prevents future information from influencing either feature construction or relevance definition, ensuring that the evaluation remains consistent with future-facing fraud detection. The resulting data partitions and query statistics are summarized in Table 8.

Table 8. Chronological data partition and operational query statistics.

Split	Transactions	Fraud cases	Fraud rate (%)	Period	Query windows
Training	1,037,340	5,968	0.5753	2019-01-01 to 2020-03-06	10,304
Validation	259,335	1,538	0.5931	2020-03-06 to 2020-06-21	2,574
Future holdout test	555,719	2,145	0.3860	2020-06-21 to 2020-12-31	4,644

The proposed formulation represents fraud severity using amount-aware graded relevance, where transaction amount serves as an operationally interpretable proxy for financial impact. Unlike conventional binary relevance, which assigns identical importance to all fraudulent transactions, the proposed design differentiates fraud cases according to transaction-amount quartiles computed from the training subset. Because the public dataset does not provide institution-specific loss estimates or investigation cost matrices, quartile-based grading offers a transparent, reproducible, and leakage-safe mechanism for representing investigation priority. The resulting relevance definitions are summarized in Table 9.

Table 9. Amount-aware graded relevance definitions.

Relevance	Condition	Operational interpretation
0	Non-fraud transaction	Not relevant for fraud investigation
1	Fraud amount ≤ 243.03	Low-severity fraud
2	243.03 < Fraud amount ≤ 397.84	Moderate-severity fraud
3	397.84 < Fraud amount ≤ 902.91	High-severity fraud
4	Fraud amount > 902.91	Very high-severity fraud

The effectiveness of this relevance formulation is evaluated by comparing Learning-to-Rank models trained with conventional binary relevance and the proposed amount-aware relevance. The comparison, summarized in Table 10, shows that incorporating graded relevance consistently improves ranking quality. At K = 10, PureLTR_amount increases NDCG from 0.201074 to 0.205367, corresponding to an absolute improvement of 0.004293 (2.14%). A similar improvement is observed at K = 50, where NDCG increases from 0.202072 to 0.206278, an absolute gain of 0.004206 (2.08%). These improvements indicate that incorporating financial severity into the relevance definition enables the ranking model to produce better top-rank ordering than conventional binary relevance.

The improvement is most apparent in NDCG@K, which directly reflects the quality of severity-aware ranking. By contrast, FraudAmountRecall@K and MeanCapturedFraudAmount@K remain relatively similar among the strongest models, particularly under larger investigation budgets. This observation suggests that the principal benefit of the proposed relevance formulation lies in improving the ordering of financially important fraud cases

rather than substantially increasing the total amount of fraud captured. Consequently, the proposed formulation primarily enhances which transactions appear first in the investigation queue rather than how many fraudulent transactions are ultimately retrieved.

Table 10. All-query ranking performance under binary and amount-aware relevance.

Model	K	NDCG@K	Recall@K	FraudAmountRecall@K	MeanCapturedFraudAmount@K
PureLTR_binary	10	0.201074	0.213677	0.216576	242.924457
PureLTR_amount	10	**0.205367**	0.213252	0.216796	242.717326
HybridLTR_amount	10	0.205217	**0.214349**	**0.217380**	**242.955080**
XGBClassifier	10	0.199055	0.213555	0.216629	242.435327
PureLTR_binary	50	0.202072	0.220295	0.220301	243.868374
PureLTR_amount	50	**0.206278**	0.220162	0.220234	243.846363
HybridLTR_amount	50	0.206211	0.221102	0.221163	243.997857
XGBClassifier	50	0.200352	**0.221469**	**0.221383**	**244.004440**

An additional observation is that the difference between HybridLTR_amount and PureLTR_amount is very small. This finding indicates that the auxiliary XGBoost-derived risk score provides only limited refinement to the final ranking, whereas the principal performance gain originates from the proposed amount-aware relevance formulation itself. This interpretation establishes the foundation for the subsequent analyses, which further examine the operational implications, the contribution of the auxiliary risk signal, and the robustness of the proposed ranking framework.

4.2. Operational Prioritization Performance under Limited Review Budget

Building upon the formulation introduced in the previous subsection, the following experiments evaluate whether the proposed ranking framework improves transaction prioritization under realistic investigation constraints. Since fraud analysts typically review only a limited number of transactions within each operational time window, the effectiveness of the proposed approach should be assessed using ranking-oriented and budget-oriented measures rather than conventional classification metrics alone. Accordingly, NDCG@K is adopted as the primary evaluation metric, while Recall@K, FraudAmountRecall@K, MeanCapturedFraudAmount@K, and Precision@K provide complementary perspectives on fraud coverage, financial impact, and investigation efficiency under fixed review budgets. All experiments employ the all-query evaluation protocol to reflect continuous fraud-monitoring environments in which ranking must be performed for every operational query, regardless of whether fraudulent transactions are present.

The overall operational performance is summarized in Table 11. Across both investigation budgets, the Learning-to-Rank models consistently achieve higher ranking quality than the pointwise classification baseline. At K = 10, HybridLTR_amount obtains an NDCG@10 of 0.205217, compared with 0.199055 for XGBClassifier. A similar pattern is observed at K = 50, where HybridLTR_amount achieves 0.206211, exceeding the 0.200352 obtained by the classifier baseline. These results indicate that optimizing the learning objective for ranking rather than classification produces more effective ordering of investigation-relevant transactions.

A more detailed comparison reveals two consistent patterns. First, PureLTR_amount and HybridLTR_amount occupy the highest positions in terms of NDCG@K, followed by ShuffledHybridControl, PureLTR_binary, and XGBClassifier. This ordering suggests that incorporating graded relevance contributes more substantially to ranking quality than simply replacing the classifier with a Learning-to-Rank algorithm. Second, the close performance between PureLTR_amount and HybridLTR_amount indicates that the principal improvement originates from the proposed relevance formulation, whereas the auxiliary classifier-derived score provides only limited additional refinement. These observations are consistent with the formulation-oriented interpretation established in Section 4.1.

The statistical significance of these improvements is evaluated using paired block-bootstrap resampling, and the results are presented in Table 12. Compared with XGBClassifier,

HybridLTR_amount significantly improves NDCG@K at both investigation budgets. At K = 10, the average improvement is +0.006162 (95% CI: 0.004552–0.007930, $p < 0.001$), while at K = 50, the improvement remains significant at +0.005860 (95% CI: 0.004326–0.007661, $p < 0.001$). Likewise, comparisons with PureLTR_binary demonstrate statistically significant gains at both budgets, indicating that the proposed amount-aware relevance formulation consistently produces better top-rank ordering than conventional binary relevance.

Table 11. All-query operational ranking performance under limited investigation budgets.

Model	K	NDCG@K	Recall@K	FraudAmountRecall@K	MeanCapturedFraudAmount@K	Precision@K
PureLTR_amount	10	**0.205367**	0.213252	0.216796	242.717326	0.043712
HybridLTR_amount	10	0.205217	**0.214349**	**0.217380**	**242.955080**	0.043885
ShuffledHybridControl	10	0.204466	0.213094	0.216801	242.756242	0.043626
PureLTR_binary	10	0.201074	0.213677	0.216576	242.924457	0.043885
XGBClassifier	10	0.199055	0.213555	0.216629	242.435327	0.043755
AmountScore	10	0.143180	0.164100	0.181988	239.298988	0.034281
Random	10	0.011540	0.022762	0.023072	20.314272	0.004414
PureLTR_amount	50	**0.206278**	0.220162	0.220234	243.846363	0.009211
HybridLTR_amount	50	0.206211	0.221102	0.221163	243.997857	0.009241
ShuffledHybridControl	50	0.205264	0.219714	0.219852	243.782905	0.009189
PureLTR_binary	50	0.202072	0.220295	0.220301	243.868374	0.009211
XGBClassifier	50	0.200352	**0.221469**	**0.221383**	**244.004440**	**0.009254**
AmountScore	50	0.145633	0.179576	0.193241	242.687168	0.007548
Random	50	0.032900	0.116276	0.115242	109.326583	0.004525

Table 12. Paired block-bootstrap comparison of the principal operational ranking metrics.

Comparison	K	Metric	Δ Mean	95% Confidence Interval	*p*-value	Interpretation
HybridLTR_amount − XGBClassifier	10	NDCG@K	+0.006162	[0.004552, 0.007930]	<0.001	Significant improvement in ranking quality
HybridLTR_amount − PureLTR_binary	10	NDCG@K	+0.004144	[0.002785, 0.005755]	<0.001	Significant improvement over binary relevance
HybridLTR_amount − PureLTR_amount	10	NDCG@K	−0.000150	[−0.001234, 0.000946]	0.715	Not significant
HybridLTR_amount − XGBClassifier	10	FraudAmountRecall@K	+0.000751	[−0.000459, 0.001845]	0.180	Not significant
HybridLTR_amount − PureLTR_binary	10	FraudAmountRecall@K	+0.000804	[−0.000161, 0.001913]	0.105	Not significant
HybridLTR_amount − XGBClassifier	50	NDCG@K	+0.005860	[0.004326, 0.007661]	<0.001	Significant improvement in ranking quality
HybridLTR_amount − PureLTR_binary	50	NDCG@K	+0.004140	[0.002869, 0.005714]	<0.001	Significant improvement over binary relevance
HybridLTR_amount − PureLTR_amount	50	NDCG@K	−0.000066	[−0.001091, 0.000959]	0.815	Not significant
HybridLTR_amount − XGBClassifier	50	FraudAmountRecall@K	−0.000220	[−0.000654, 0.000000]	0.080	Not significant

An important observation is that the greatest improvement consistently appears in NDCG@K, whereas the differences in FraudAmountRecall@K are comparatively small among the strongest models. This behavior is expected because the proposed framework is

optimized to improve the ordering of transactions according to graded relevance rather than to maximize the total amount of fraud captured within a fixed review budget. Consequently, two models may retrieve similar amounts of fraudulent transactions while differing substantially in how effectively high-severity cases are positioned near the top of the investigation queue. The statistically significant gains in NDCG@K, together with the consistently small differences between HybridLTR_amount and PureLTR_amount, therefore reinforce the conclusion that the primary empirical advantage arises from the proposed amount-aware ranking formulation rather than from the auxiliary classifier-derived signal. This interpretation naturally motivates the following subsection, which examines the specific contribution of the hybrid XGBoost risk score to the overall framework.

4.3. Role of the Hybrid XGBoost Signal

The previous subsection demonstrated that the proposed amount-aware ranking formulation consistently improves operational ranking quality. This subsection further examines whether the observed improvement is primarily attributable to the proposed formulation itself or to the auxiliary XGBoost-derived risk signal incorporated into the hybrid framework. Rather than serving as the final prediction, the classifier-derived score is used as an additional feature to complement the engineered transaction representation, while the final transaction ordering remains determined by the LambdaMART ranker using amount-aware graded relevance. This design enables the auxiliary classifier to provide complementary risk information without altering the underlying ranking objective.

To ensure that the auxiliary score remains free from future information, it is generated using a chronological out-of-fold strategy. The training period is divided into three expanding folds, where each model is trained only on earlier observations and evaluated on the immediately following temporal segment. As summarized in Table 14, the optimal boosting iteration remains highly consistent across all folds, ranging from 219 to 224, indicating stable learning behaviour throughout the chronological training process. Together with the strong discrimination performance reported previously for the future holdout set, these results support the use of the XGBoost output as an informative auxiliary feature rather than as the final operational decision.

Table 14. Chronological out-of-fold performance of the auxiliary XGBoost classifier.

Fold	Training size	Validation size	Best iteration	Best score
1	259,335	259,335	224	0.992924
2	518,670	259,335	223	0.995707
3	778,005	259,335	219	0.997000

The practical contribution of the auxiliary score is evaluated by comparing HybridLTR_amount, PureLTR_amount, and ShuffledHybridControl. As reported in Table 11, the difference between PureLTR_amount and HybridLTR_amount is negligible in terms of NDCG@K. At K = 10, the corresponding NDCG values are 0.205367 and 0.205217, while at K = 50 they are 0.206278 and 0.206211, respectively. These results indicate that introducing the auxiliary classifier-derived score does not materially alter the ordering quality achieved by the amount-aware ranking formulation.

Although the improvement in NDCG@K is marginal, the auxiliary score consistently produces slightly higher Recall@K, FraudAmountRecall@K, and MeanCapturedFraudAmount@K than the pure ranker. This pattern suggests that the classifier-derived score provides complementary information that helps retrieve additional fraud cases within the limited review budget without substantially changing the graded ordering learned by the ranking model. Consequently, the hybrid component functions primarily as a refinement mechanism for fraud capture rather than as the principal contributor to ranking quality.

The role of the auxiliary signal is further examined using ShuffledHybridControl, in which the classifier-derived score is randomly permuted before being supplied to the ranker. If the auxiliary score contained no meaningful predictive information, the shuffled and original hybrid models would be expected to produce similar performance. Instead, HybridLTR_amount consistently achieves higher Recall@K, FraudAmountRecall@K, and MeanCapturedFraudAmount@K, particularly under the larger investigation budget (K = 50).

This comparison confirms that the original classifier-derived score contains useful fraud-risk information, even though its influence on ranking quality remains relatively small.

The statistical comparisons presented in Table 15 further clarify this interpretation. The differences in NDCG@K between HybridLTR_amount and PureLTR_amount are not statistically significant at either investigation budget ($p = 0.715$ at $K = 10$ and $p = 0.815$ at $K = 50$). In contrast, under $K = 50$, the hybrid model produces statistically significant improvements in Recall@K and FraudAmountRecall@K relative to the pure ranker, while also outperforming the shuffled-score control on both metrics. These findings indicate that the auxiliary score contributes primarily to fraud retrieval rather than to the graded ordering itself.

Table 15. Paired block-bootstrap comparison of the hybrid ranking variants.

Comparison	K	Metric	Δ Mean	95% Confidence Interval	p-value	Interpretation
HybridLTR_amount − PureLTR_amount	10	NDCG@K	−0.000150	[−0.001234, 0.000946]	0.715	Not significant
	10	Recall@K	+0.001097	[−0.000297, 0.002641]	0.145	Not significant
	10	FraudAmountRecall@K	+0.000584	[−0.000687, 0.001739]	0.365	Not significant
HybridLTR_amount − ShuffledHybridControl	10	NDCG@K	+0.000751	[−0.000431, 0.001864]	0.195	Not significant
	10	Recall@K	+0.001255	[−0.000122, 0.002595]	0.060	Marginal, not significant
	10	FraudAmountRecall@K	+0.000579	[−0.000519, 0.001613]	0.295	Not significant
HybridLTR_amount − PureLTR_amount	50	NDCG@K	−0.000066	[−0.001091, 0.000959]	0.815	Not significant
	50	Recall@K	+0.000940	[0.000187, 0.001760]	<0.001	Significant improvement in fraud retrieval
	50	FraudAmountRecall@K	+0.000929	[0.000186, 0.001722]	<0.001	Significant improvement in amount-weighted fraud retrieval
HybridLTR_amount − ShuffledHybridControl	50	NDCG@K	+0.000947	[−0.000098, 0.001964]	0.080	Marginal, not significant
	50	Recall@K	+0.001389	[0.000545, 0.002411]	<0.001	Significant improvement over shuffled auxiliary score
	50	FraudAmountRecall@K	+0.001311	[0.000421, 0.002312]	<0.001	Significant improvement over shuffled auxiliary score

4.4. Investigation-Budget and Risk-Concentration Implications

Having established that the proposed amount-aware ranking formulation improves ordering quality and that the auxiliary XGBoost signal provides only incremental refinement, the following analyses examine the practical implications of the proposed framework under realistic investigation constraints. In operational fraud detection, analysts are typically able to inspect only a limited number of transactions within each monitoring period. Consequently, the effectiveness of a ranking model depends not only on its overall predictive performance but also on how efficiently it concentrates high-risk transactions within a restricted review budget. The investigation-budget analysis therefore evaluates the proposed framework across $K = 5$, 10, 20, 50, and 100, where each value represents the maximum number of transactions reviewed in each operational query window.

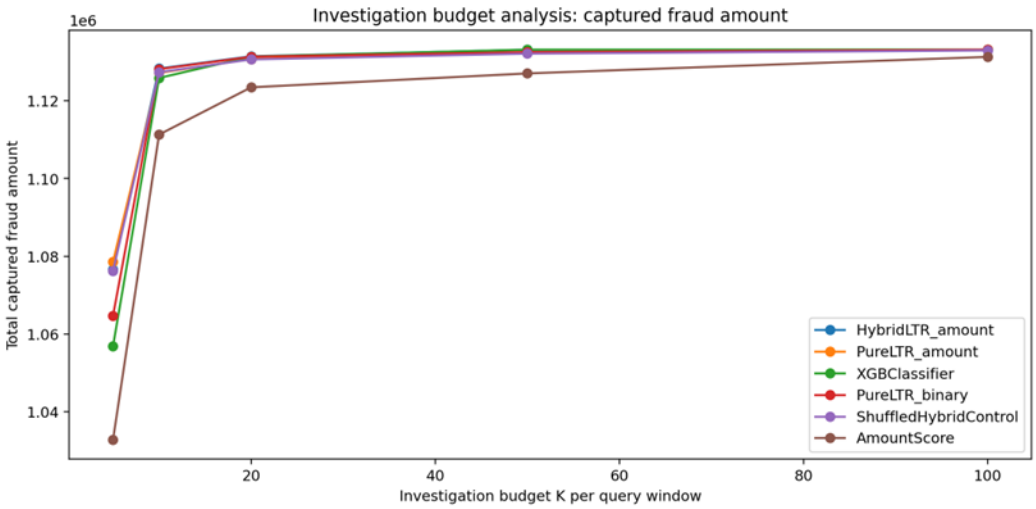

Figure 2. Investigation-budget analysis based on total captured fraud amount across K = 5, 10, 20, 50, and 100.

As illustrated in Figure 2, the amount of recovered fraud increases rapidly when the investigation budget expands from K = 5 to K = 10, whereas the incremental benefit gradually diminishes as larger review budgets become available. This pattern indicates that the proposed ranking framework delivers its greatest practical value under strict investigation constraints, where correctly prioritizing a small number of transactions has the largest operational impact. The amount-aware ranking variants consistently outperform the simple AmountScore baseline and remain competitive with the pointwise XGBClassifier. Furthermore, both HybridLTR_amount and PureLTR_amount approach a performance plateau near K = 20, suggesting that most recoverable fraud has already been concentrated within a relatively small review list. The operational interpretation of these investigation-budget regimes is summarized in Table 16.

Table 16. Operational interpretation of different investigation-budget regimes.

Investigation budget	Empirical observation	Operational implication
K = 5	Largest separation between competing models	Ranking quality is critical because only a few transactions can be investigated.
K = 10	Sharp increase in captured fraud amount	Small increases in analyst capacity yield substantial operational benefit.
K = 20	Performance begins to stabilize.	Most financially important fraud cases have already entered the review queue.
K = 50	Differences among stronger models become smaller	Additional review capacity provides diminishing operational returns.
K = 100	Performance approaches saturation	Relaxed investigation budget with limited additional practical benefit.

The investigation-budget analysis demonstrates that the proposed formulation is particularly effective when review resources are scarce. However, prioritization performance is meaningful only if the ranking model is capable of concentrating fraud within a relatively small portion of the ranked transactions. To examine this property, the future-holdout transactions are sorted according to the HybridLTR_amount ranking score and divided into ten equal-sized risk-score deciles, where D1 represents the highest-risk transactions and D10 the lowest.

The resulting fraud distribution, summarized in Table 17, shows an exceptionally strong concentration of fraud within the highest-score segment. The first decile (D1) contains 2,134 of the 2,145 fraudulent transactions in the future-holdout set, corresponding to 99.49% of all observed fraud cases. Its fraud rate reaches 3.8401%, whereas the remaining deciles contain very few fraudulent transactions and several lower-score groups contain none at all. These

results indicate that the proposed ranking framework successfully compresses the vast majority of operationally relevant cases into a compact investigation region.

Table 17. Fraud concentration across HybridLTR_amount ranking-score deciles on the future-holdout test set.

Risk-score decile	Transactions	Fraud cases	Fraud rate (%)	Fraud amount
D1	55,572	2,134	3.8401	1,132,345.04
D2	55,572	6	0.0108	561.07
D3	55,572	2	0.0036	243.26
D4	55,572	1	0.0018	1.99
D5	55,572	1	0.0018	5.26
D6	55,572	0	0.0000	0.00
D7	55,572	0	0.0000	0.00
D8	55,572	0	0.0000	0.00
D9	55,572	0	0.0000	0.00
D10	55,571	1	0.0018	4.11
Total	555,719	2,145	0.3860	1,133,160.73

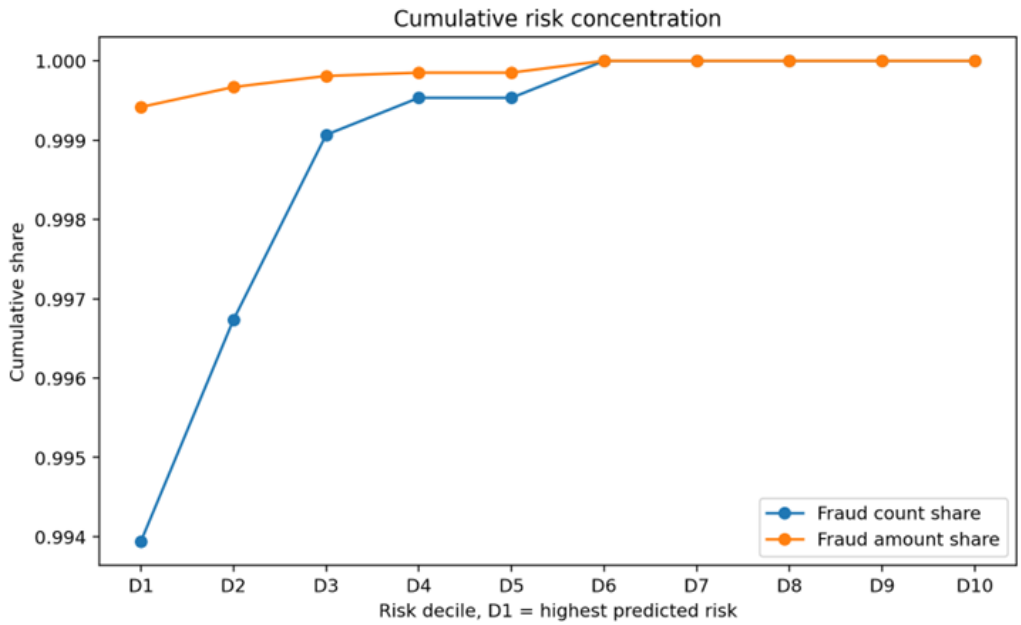

Figure 3. Cumulative fraud-count and fraud-amount concentration across HybridLTR_amount ranking-score deciles.

The concentration pattern becomes even more evident when cumulative fraud counts and fraud amounts are examined. As illustrated in Figure 3, the highest-score decile alone contains 99.49% of all fraud cases and 99.93% of the total fraudulent transaction amount observed in the future-holdout period. The cumulative curves therefore confirm that the proposed framework concentrates nearly all operationally significant fraud into a very small proportion of ranked transactions, substantially reducing the effective search space for fraud analysts.

To complement the aggregate analyses, a transaction-level inspection of the highest-ranked cases is provided in Table 18. Rather than serving as quantitative evidence, this analysis illustrates the practical characteristics of the generated investigation queue. All 100 highest-ranked transactions correspond to confirmed fraud cases, with a cumulative fraud amount of 104,293.13, while shopping_net emerges as the dominant transaction context. This qualitative example demonstrates that the proposed ranking framework produces an operationally interpretable review list composed primarily of financially significant and highly suspicious transactions.

Table 18. Qualitative characteristics of the highest-ranked investigation queue.

Indicator	Observation
Transactions reviewed	100
Fraudulent transactions	100
Total fraud amount	104,293.13
Dominant transaction context	`shopping_net`
Overall pattern	High-value and operationally suspicious transactions

Taken together, the investigation-budget analysis, risk-concentration evaluation, cumulative-decile visualization, and transaction-level inspection provide complementary evidence of the proposed framework's operational effectiveness. The budget analysis demonstrates that most recoverable fraud can be identified using relatively small review capacities, whereas the decile analyses show that nearly all fraud cases and fraudulent transaction amounts are concentrated within the highest-ranked transactions. Rather than merely improving predictive performance, the proposed amount-aware ranking formulation produces a compact, high-yield investigation queue that better aligns model outputs with the practical workflow of fraud analysts operating under limited investigation resources.

4.5. Robustness and Additional Analysis

4.5.1. Temporal Robustness and Generalization

The effectiveness of the proposed amount-aware ranking formulation has been demonstrated under the primary future-holdout evaluation. To further assess its practical applicability, additional experiments are conducted to examine whether the framework remains reliable under different temporal conditions and operational configurations. Specifically, this subsection evaluates temporal robustness across monthly test periods, analyzes feature distribution stability, investigates the influence of different query-window sizes, and validates the formulation using multiple rolling-origin chronological splits. Together, these analyses provide evidence of the proposed framework's ability to generalize beyond a single evaluation setting.

Table 19. Monthly robustness of the proposed HybridLTR_amount model under the all-query evaluation protocol (K = 10).

Time block	Fraud cases	Fraud rate (%)	Fraud amount	NDCG@10	Recall@10	FraudAmount Recall@10
2020-06	133	0.4425	73,274.93	0.238917	0.246223	0.253662
2020-07	321	0.3739	158,669.49	0.193284	0.202269	0.204948
2020-08	415	0.4676	208,785.43	0.219539	0.234124	0.240224
2020-09	340	0.4890	202,700.99	0.230384	0.238199	0.239814
2020-10	384	0.5537	195,572.97	0.218707	0.227281	0.229023
2020-11	294	0.4048	153,182.19	0.207967	0.213495	0.215203
2020-12	258	0.1849	141,138.68	0.151996	0.161703	0.164602

The monthly robustness analysis evaluates the proposed framework throughout the future-holdout period to determine whether ranking performance remains consistent under changing fraud characteristics. As summarized in Table 19, the monthly evaluation exhibits natural performance variation across different operational periods. The highest performance is observed in June 2020, where HybridLTR_amount achieves an NDCG@10 of 0.238917, together with Recall@10 and FraudAmountRecall@10 values of 0.246223 and 0.253662, respectively. By comparison, the lowest performance occurs in December 2020, where NDCG@10 decreases to 0.151996 while the fraud rate simultaneously declines to 0.1849%. These observations suggest that the difficulty of fraud prioritization varies naturally with changing fraud prevalence and transaction characteristics rather than remaining constant throughout the evaluation period.

Although the monthly metrics fluctuate, this variation should not be interpreted as instability of the proposed formulation. Instead, it reflects the evolving nature of fraud detection, where fraud prevalence and transaction distributions change over time. To further

distinguish temporal variation from feature instability, the Population Stability Index (PSI) is examined for the principal input variables.

As presented in Table 20, the largest distribution shifts occur in the temporal variables, particularly unix_time and month, with PSI values of 12.433856 and 4.390929, respectively. These shifts are expected because the evaluation deliberately employs a chronological future-holdout protocol. In contrast, behavioral and demographic variables exhibit very small PSI values, indicating that most non-temporal characteristics remain highly stable between the training and testing periods. Consequently, the observed monthly performance variation is primarily associated with intended temporal separation rather than widespread feature drift.

Table 20. Highest Population Stability Index (PSI) values between the training and future-holdout periods.

Feature	PSI (Train vs Test)	Interpretation
unix_time	12.433856	Strong temporal separation by design
month	4.390929	Strong calendar-period shift
dayofweek	0.080150	Mild temporal-pattern shift
dow_sin	0.053901	Mild weekly-cycle shift
dow_cos	0.025474	Small weekly-cycle shift
day	0.012606	Small day-level shift
age	0.009342	Negligible demographic shift
cc_num_freq	0.000289	Negligible behavioral-frequency shift
city_freq	0.000220	Negligible behavioral-frequency shift
merchant_freq	0.000100	Negligible behavioral-frequency shift

Beyond temporal robustness, the proposed formulation is evaluated under different operational query granularities to determine whether its effectiveness depends heavily on a specific query-window configuration. The results summarized in Table 21 show that larger query windows consistently produce higher ranking metrics because more transactions become available within each operational query. For example, NDCG@10 increases from 0.083964 for a 15-minute window to 0.205292 for the 1-hour configuration, reaching 0.519709 for a 6-hour window. Similar trends are observed for FraudAmountRecall@K.

This improvement, however, should not be interpreted as evidence that larger windows are universally preferable. Increasing the query duration naturally enlarges the candidate transaction pool, thereby making the ranking task easier while reducing operational responsiveness. The one-hour configuration adopted throughout this study therefore represents a practical compromise between investigation latency and sufficient candidate density for effective ranking.

Table 21. Sensitivity analysis of different operational query-window durations.

Query Window	Test Queries	NDCG@10	Recall@10	FraudAmount Recall@10	NDCG@50	Recall@50	FraudAmount Recall@50
15 min	18,576	0.083964	0.087595	0.087639	0.084120	0.088232	0.088232
30 min	9,288	0.137062	0.144031	0.144453	0.137439	0.145887	0.145887
1 hour	**4,644**	**0.205292**	**0.214601**	**0.217596**	**0.206226**	**0.221102**	**0.221163**
2 hours	2,322	0.285308	0.295090	0.304974	0.288143	0.315157	0.315774
6 hours	774	0.519709	0.525257	0.567776	0.526277	0.581003	0.588206

To further evaluate temporal generalization, rolling-origin validation is performed using three independent chronological train-test partitions. Unlike the single future-holdout evaluation, this experiment examines whether the proposed formulation remains effective across multiple temporal splits.

The results, presented in Table 22, demonstrate consistent ranking performance throughout the three rolling folds. NDCG@10 ranges from 0.219384 to 0.254698, while NDCG@50 varies between 0.221201 and 0.256200. Likewise, FraudAmountRecall@10 remains consistently high, ranging from 0.231640 to 0.274485, despite differences in fraud

prevalence across evaluation periods. These findings indicate that the proposed amount-aware ranking formulation generalizes well under different chronological partitions rather than depending on a single train-test split.

Table 22. Rolling-origin validation of the proposed amount-aware operational ranking framework.

Fold	Train end fraction	Test period	K	NDCG@K	Recall@K	FraudAmount Recall@K	Test fraud rate (%)
1	0.45	2019-10-31 to 2019-12-14	10	0.249831	0.260692	0.266148	0.4990
1	0.45	2019-10-31 to 2019-12-14	50	0.250516	0.267833	0.268677	0.4990
2	0.65	2020-01-28 to 2020-04-03	10	0.219384	0.227644	0.231640	0.6555
2	0.65	2020-01-28 to 2020-04-03	50	0.221201	0.238185	0.238608	0.6555
3	0.85	2020-05-29 to 2020-06-21	10	0.254698	0.270827	0.274485	0.5691
3	0.85	2020-05-29 to 2020-06-21	50	0.256200	0.280048	0.280275	0.5691

Taken together, the monthly robustness analysis, feature-stability assessment, query-window sensitivity study, and rolling-origin validation provide consistent evidence that the proposed framework remains reliable under realistic future-oriented evaluation settings. Although absolute performance naturally varies with changing fraud prevalence and temporal transaction characteristics, the observed variation reflects the evolving operational environment rather than instability of the proposed formulation. Collectively, these results support the generalizability of the amount-aware operational ranking framework across different temporal conditions while reinforcing its suitability for practical fraud-investigation scenarios.

4.5.2. Component Contribution Analysis

Although the previous subsection demonstrates that the proposed framework generalizes well under different temporal conditions, temporal robustness alone does not explain which components are primarily responsible for the observed performance improvement. The following analyses therefore investigate the contribution of the principal design elements through ablation experiments and statistical significance testing. In particular, the experiments examine whether the performance gain originates from the proposed amount-aware relevance formulation, the auxiliary XGBoost-derived risk score, or the major feature groups incorporated into the ranking model.

The ablation results are summarized in Table 23. Across both investigation budgets, replacing the proposed amount-aware relevance with conventional binary relevance produces the largest reduction in ranking quality. At $K = 10$, the BinaryRelevance variant decreases NDCG@10 by 0.006682, while at $K = 50$ the reduction remains substantial at 0.006440. These degradations are considerably larger than those observed when removing the auxiliary hybrid score or excluding individual feature groups. This pattern indicates that the empirical improvement is primarily attributable to the proposed relevance formulation rather than to the additional classifier-derived signal or any individual feature subset.

A more detailed examination reveals three consistent observations. First, removing the amount-aware relevance produces the largest degradation in NDCG@K, confirming that the learning objective itself constitutes the principal contribution of the proposed framework. Second, excluding temporal or amount-related features also reduces ranking quality, although the magnitude of these changes is substantially smaller. Third, removing the auxiliary XGBoost-derived score has only a negligible effect on NDCG@K, indicating that the hybrid component functions primarily as a complementary enhancement rather than as the dominant source of ranking improvement. These findings are consistent with the formulation-oriented interpretation established throughout the preceding sections.

The statistical comparisons presented in Table 24 further reinforce these observations. Replacing amount-aware relevance with binary relevance results in statistically significant

reductions in NDCG@K at both investigation budgets ($p < 0.001$). Likewise, removing temporal or amount-related features also produces significant performance degradation, indicating that these feature groups contribute meaningfully to the final ranking model. In contrast, removing the auxiliary hybrid score does not produce a statistically significant reduction in NDCG@K at either $K = 10$ ($p = 0.715$) or $K = 50$ ($p = 0.815$). Similarly, randomly shuffling the hybrid score weakens ranking performance only marginally and does not significantly affect the primary ranking metric.

Table 23. Ablation analysis for identifying the principal source of performance improvement.

Ablation Variant	K	NDCG@K	Δ NDCG	Recall@K	Δ Recall	FraudAmount Recall@K	Δ FraudAmount Recall
ProposedFull	10	0.205217	0.000000	0.214349	0.000000	0.217380	0.000000
NoHybridScore	10	0.205367	+0.000150	0.213252	−0.001097	0.216796	−0.000584
BinaryRelevance	10	0.198535	−0.006682	0.212859	−0.001491	0.215987	−0.001393
NoTemporalFeatures	10	0.203651	−0.001566	0.213233	−0.001117	0.216576	−0.000803
NoAmountFeatures	10	0.203194	−0.002023	0.212678	−0.001671	0.215954	−0.001426
ShuffledHybridScore	10	0.204466	−0.000751	0.213094	−0.001255	0.216801	−0.000579
ProposedFull	50	0.206211	0.000000	0.221102	0.000000	0.221163	0.000000
NoHybridScore	50	0.206278	+0.000066	0.220162	−0.000940	0.220234	−0.000929
BinaryRelevance	50	0.199771	−0.006440	0.220457	−0.000646	0.220517	−0.000646
NoTemporalFeatures	50	0.204831	−0.001380	0.220887	−0.000215	0.220948	−0.000215
NoAmountFeatures	50	0.204592	−0.001619	0.221038	−0.000065	0.220952	−0.000211
ShuffledHybridScore	50	0.205264	−0.000947	0.219714	−0.001389	0.219852	−0.001311

Table 24. Paired block-bootstrap statistical comparison of key ablation variants based on NDCG.

Comparison	K	Δ NDCG	95% Confidence Interval	p-value	Interpretation
ProposedFull − NoHybridScore	10	−0.000150	[−0.001234, 0.000946]	0.715	No significant loss without the auxiliary hybrid score
ProposedFull − BinaryRelevance	10	+0.006682	[0.004978, 0.008315]	<0.001	Significant degradation after replacing amount-aware relevance
ProposedFull − NoTemporalFeatures	10	+0.001566	[0.000662, 0.002426]	<0.001	Temporal features contribute significantly
ProposedFull − NoAmountFeatures	10	+0.002023	[0.000931, 0.003227]	<0.001	Amount-related features contribute significantly
ProposedFull − ShuffledHybridScore	10	+0.000751	[−0.000431, 0.001864]	0.195	Marginal degradation after score randomization
ProposedFull − NoHybridScore	50	−0.000066	[−0.001091, 0.000959]	0.815	No significant loss without the auxiliary hybrid score
ProposedFull − BinaryRelevance	50	+0.006440	[0.004865, 0.008044]	<0.001	Significant degradation after replacing amount-aware relevance
ProposedFull − NoTemporalFeatures	50	+0.001380	[0.000577, 0.002226]	<0.001	Temporal features contribute significantly
ProposedFull − NoAmountFeatures	50	+0.001619	[0.000546, 0.002740]	<0.001	Amount-related features contribute significantly
ProposedFull − ShuffledHybridScore	50	+0.000947	[−0.000098, 0.001964]	0.080	Marginal degradation after score randomization

Taken together, the ablation experiments and statistical analyses provide consistent evidence regarding the origin of the observed performance improvement. The largest and most statistically significant degradation occurs when the amount-aware relevance formulation is removed, whereas eliminating the auxiliary hybrid score produces no significant loss in ranking quality. These findings strengthen the central claim of this study by demonstrating that

the principal contribution lies in reformulating fraud detection as an amount-aware operational ranking problem rather than in introducing an additional hybrid classification component. The auxiliary XGBoost-derived score remains valuable as a complementary source of fraud-risk information, but its role is to refine fraud retrieval rather than to define the ranking objective itself. This interpretation provides a coherent explanation for the empirical observations presented throughout Sections 4.1–4.4 and establishes a clear foundation for the model interpretation presented in the following subsection.

4.5.3. Model Interpretation and Dataset-Family Positioning

Beyond demonstrating strong predictive performance, interpreting the learned ranking model provides additional insight into how the proposed framework prioritizes fraudulent transactions. Such interpretation is particularly important because the primary contribution of this study lies in the amount-aware operational ranking formulation rather than in the introduction of a new learning algorithm. Accordingly, this subsection examines the feature contributions learned by the ranking model and subsequently positions the proposed framework within the broader family of Sparkov-based fraud detection studies.

The feature-importance analysis is summarized in Table 25. The learned ranking model assigns the highest importance to the transaction amount (amt) and its logarithmic representation (log_amt), followed by the auxiliary xgb_oof_risk_score. This ordering is consistent with the proposed formulation, in which transaction amount directly determines the graded relevance used for ranking, while the classifier-derived score serves as complementary risk information. The remaining influential variables consist primarily of behavioral-frequency features, temporal attributes, and encoded categorical variables, indicating that the final ranking is produced through a combination of financial severity, transaction behavior, and temporal context rather than by relying on a single predictor.

Table 25. Top feature gains learned by the HybridLTR_amount ranking model.

Rank	Feature	Gain	Interpretation
1	amt	198.317352	Direct transaction-amount signal
2	log_amt	98.527054	Scale-adjusted amount representation
3	xgb_oof_risk_score	55.868195	Auxiliary classifier-derived risk signal
4	category_freq	46.690842	Behavioral frequency information
5	unix_time	18.818117	Temporal information
6	category_hash	15.113388	Encoded transaction-category identifier
7	hour	10.219227	Time-of-day behavioral pattern
8	merchant_freq	9.426815	Merchant-level transaction frequency
9	hour_cos	8.903326	Cyclical temporal pattern
10	gender_hash	7.344675	Encoded contextual information

The feature ranking provides additional evidence supporting the conclusions drawn from the ablation study. Specifically, the dominance of amt and log_amt is consistent with the observation that replacing the amount-aware relevance formulation with binary relevance produces the largest degradation in NDCG@K. Likewise, the position of xgb_oof_risk_score as the third most influential feature agrees with the earlier finding that the auxiliary classifier contributes useful fraud-risk information while remaining secondary to the amount-aware formulation. Together, the feature-importance and ablation analyses present a consistent interpretation of how the proposed ranking framework produces its operational prioritization.

To further contextualize the proposed framework, Table 26 compares this study with representative fraud detection approaches evaluated on the Sparkov family of simulated credit-card transaction datasets. Rather than providing a direct state-of-the-art comparison, this analysis highlights the different research objectives addressed by previous studies and by the proposed framework. The comparison illustrates that previous Sparkov-family studies have consistently achieved excellent fraud-classification performance using conventional machine-learning algorithms, AutoML frameworks, deep-learning architectures, and graph neural networks. These studies demonstrate that the dataset family contains rich predictive information and has been extensively investigated from a classification perspective. However, their

primary objective is to determine whether individual transactions are fraudulent, with evaluation typically based on AUC-ROC, F1-score, precision, or recall.

Table 26. Dataset-family positioning of the proposed framework relative to representative Sparkov-based fraud detection studies.

Study	Dataset relation	Main method	Primary objective	Reported evaluation focus	Relation to this study
[27]	Includes Sparkov, a simulated credit-card transaction dataset within the Sparkov/Kaggle family	Random Forest, CatBoost, LightGBM, MLP, and AutoML frameworks	Binary fraud classification	AUC-ROC (CatBoost 0.995; AFD OFI 0.998; AutoGluon and H2O 0.997; Auto-sklearn 0.995)	Provides a strong classification benchmark but does not consider operational ranking
[28]	Uses Sparkov-generated transaction data with numerical and textual attributes	BERT-LSTM hybrid network	Fraud classification using sequential and textual information	Accuracy, Precision, Recall, and F1-score	Demonstrates deep-learning classification but not investigation prioritization
[29]	Uses a large-scale Sparkov-based customer-merchant transaction graph	Encoder-decoder Graph Neural Network	Graph-based fraud classification	Precision, Recall, F1-score, and AUC-ROC	Provides graph-based contextual modeling without amount-aware ranking
Proposed framework	Uses the Kartik Shenoy dataset [30], where fraudTrain.csv is chronologically divided into training and validation sets, and fraudTest.csv is retained as the future holdout	Amount-aware Learning-to-Rank with auxiliary XGBoost-derived risk score	Operational fraud prioritization	NDCG@K, FraudAmountRecall@K, MeanCapturedFraudAmount@K, and risk concentration	Introduces a formulation-oriented approach for severity-aware investigation prioritization

By contrast, the present study addresses a different operational problem. Rather than maximizing binary classification performance, the proposed framework focuses on determining which fraudulent transactions should be investigated first when analyst resources are limited and fraudulent transactions differ substantially in financial severity. Consequently, the comparison presented in Table 26 should be interpreted as dataset-family positioning rather than as a direct state-of-the-art performance comparison, since the learning objectives and evaluation criteria differ fundamentally.

Taken together, the feature-importance analysis and the dataset-family comparison reinforce the central contribution of this study. The learned ranking behavior is fully consistent with the proposed amount-aware relevance formulation, while the comparison with previous Sparkov-based studies highlights that the present work addresses a complementary research direction centered on operational investigation prioritization rather than conventional binary fraud classification. Combined with the robustness analyses and component-level evaluation presented in the preceding subsections, these findings provide converging evidence that reformulating fraud detection as an amount-aware operational ranking problem offers a practically meaningful framework for supporting fraud analysts operating under limited investigation budgets.

5. Conclusions

This study reformulated transaction fraud detection as an amount-aware operational ranking problem rather than a conventional binary classification task. Motivated by the practical constraints of fraud investigation, the proposed framework assigns graded relevance using training-only transaction-amount quartiles and ranks transactions within operational time-window queries, thereby aligning the learning objective with severity-aware investigation prioritization. The experimental results demonstrate that the proposed formulation consistently improves operational fraud prioritization compared with conventional binary relevance. The ablation and statistical analyses further show that the observed performance gains originate primarily from the amount-aware relevance formulation rather than from the auxiliary XGBoost-derived risk score. Operational evaluations also indicate that substantial fraud amounts can be captured within relatively small investigation budgets while maintaining consistent performance across different temporal evaluation settings.

This study has several limitations. Transaction amount is used as a proxy for financial severity because the public dataset does not provide institution-specific information such as

realized financial loss, recovery outcome, or investigation cost. In addition, the evaluation is conducted on a single public dataset using LambdaMART as the ranking model, meaning that the reported findings primarily validate the proposed ranking formulation rather than a specific learning architecture. Future work may extend the proposed framework by incorporating institution-specific cost functions or adaptive relevance definitions, investigating drift-aware learning-to-rank models for evolving fraud patterns, and evaluating alternative ranking architectures under the same operational protocol using diverse real-world financial datasets. Such extensions would further strengthen the applicability of amount-aware operational ranking as a practical decision-support framework for fraud investigation under limited analyst resources.

Funding: This research was funded by Universitas Amikom Yogyakarta.

Acknowledgments: The author also acknowledges the use of AI-assisted tools for language refinement, structural editing, and support for manuscript preparation. All conceptual design, methodological decisions, experimental analysis, interpretation of results, and final manuscript responsibility remain with the author.

Conflicts of Interest: The author declares no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] R. R. Popat and J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, May 2018, pp. 1120–1125. doi: 10.1109/ICOEI.2018.8553963.

[2] K. H. Ahmed, S. Axelsson, Y. Li, and A. M. Sagheer, "A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling," *Mach. Learn. with Appl.*, vol. 20, p. 100675, Jun. 2025, doi: 10.1016/j.mlwa.2025.100675.

[3] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Syst. Appl.*, vol. 193, p. 116429, May 2022, doi: 10.1016/j.eswa.2021.116429.

[4] T. A. Gaav, H. U. Adoga, and T. Moses, "Recent Advances in Credit Card Fraud Detection: An Analytical Review of Frameworks, Methodologies, Datasets, and Challenges," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 343–369, Sep. 2025, doi: 10.62411/faith.3048-3719-251.

[5] C.-T. Chen, C. Lee, S.-H. Huang, and W.-C. Peng, "Credit Card Fraud Detection via Intelligent Sampling and Self-supervised Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–29, Apr. 2024, doi: 10.1145/3641283.

[6] H. Xu, "Trustworthy Fraud Detection under Distribution Shift: Calibrated, Cost-Optimal Ensembles with Statistical Guarantees and External Validation," in *Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing*, Nov. 2025, pp. 86–98. doi: 10.1145/3788910.3788923.

[7] T.-Y. Liu, "Learning to Rank for Information Retrieval," *Found. Trends® Inf. Retr.*, vol. 3, no. 3, pp. 225–331, Jun. 2009, doi: 10.1561/1500000016.

[8] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An Overview," 2010. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/MSR-TR-2010-82.pdf

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[10] S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," *Expert Syst. Appl.*, vol. 240, p. 122156, Apr. 2024, doi: 10.1016/j.eswa.2023.122156.

[11] D. Cheng, Y. Zou, S. Xiang, and C. Jiang, "Graph neural networks for financial fraud detection: a review," *Front. Comput. Sci.*, vol. 19, no. 9, p. 199609, Sep. 2025, doi: 10.1007/s11704-024-40474-y.

[12] Y. Chen, C. Zhao, Y. Xu, C. Nie, and Y. Zhang, "Deep learning in financial fraud detection: Innovations, challenges, and applications," *Data Sci. Manag.*, vol. 9, no. 2, p. 100162, Jun. 2026, doi: 10.1016/j.dsm.2025.08.002.

[13] I. Y. Hafez, A. Y. Hafez, A. Saleh, A. A. Abd El-Mageed, and A. A. Abohany, "A systematic review of AI-enhanced techniques in credit card fraud detection," *J. Big Data*, vol. 12, no. 1, p. 6, Jan. 2025, doi: 10.1186/s40537-024-01048-8.

[14] W. W. Li and T. Ma, "Learn to Rank Risky Investors: A Case Study of Predicting Retail Traders' Behaviour and Profitability," *ACM Trans. Inf. Syst.*, vol. 44, no. 1, pp. 1–33, Jan. 2026, doi: 10.1145/3768623.

[15] D. N. Wicaksono, D. R. I. M. Setiadi, A. Susanto, I. Harkespan, M. A. Mohamed, and A. Sambas, "Understanding Statistical and Temporal Representations for Large-Scale IoT DDoS Detection Through Ablation-Driven Analysis," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 678–698, May 2026, doi: 10.62411/jcta.16126.

[16] B. Yousefimehr and M. Ghatee, "A distribution-preserving method for resampling combined with LightGBM-LSTM for sequence-wise fraud detection in credit card transactions," *Expert Syst. Appl.*, vol. 262, p. 125661, Mar. 2025, doi: 10.1016/j.eswa.2024.125661.

[17] Y.-C. Shih, T.-S. Dai, Y.-P. Chen, Y.-W. Ti, W.-H. Wang, and Y. Kuo, "Fund transfer fraud detection: Analyzing irregular transactions and customer relationships with self-attention and graph neural networks," *Expert Syst. Appl.*, vol. 259, p. 125211, Jan. 2025, doi: 10.1016/j.eswa.2024.125211.

[18] W. Hyun, I. Lee, and B. Suh, "LEX-GNN: Label-Exploring Graph Neural Network for Accurate Fraud Detection," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, Oct. 2024, pp. 3802–3806. doi: 10.1145/3627673.3679956.
[19] J. Lin, X. Guo, Y. Zhu, S. Mitchell, E. Altman, and J. Shun, "FraudGT: A Simple, Effective, and Efficient Graph Transformer for Financial Fraud Detection," in *Proceedings of the 5th ACM International Conference on AI in Finance*, Nov. 2024, pp. 292–300. doi: 10.1145/3677052.3698648.
[20] L. Han *et al.*, "Mitigating the Tail Effect in Fraud Detection by Community Enhanced Multi-Relation Graph Neural Networks," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 4, pp. 2029–2041, Apr. 2025, doi: 10.1109/TKDE.2025.3530467.
[21] G. Tong, J. Qian, and J. Shen, "Adaptive metagraph neural network assisted by metagraph search for financial fraud detection," *Eng. Appl. Artif. Intell.*, vol. 153, p. 110807, Aug. 2025, doi: 10.1016/j.engappai.2025.110807.
[22] T. John Berkmans and S. Karthick, "Anomaly detection in online credit card data using optimized multi-view heterogeneous graph neural networks," *Knowledge-Based Syst.*, vol. 324, p. 113767, Aug. 2025, doi: 10.1016/j.knosys.2025.113767.
[23] J. Qian and G. Tong, "Metapath-guided graph neural networks for financial fraud detection," *Comput. Electr. Eng.*, vol. 126, p. 110428, Aug. 2025, doi: 10.1016/j.compeleceng.2025.110428.
[24] X. Wang, J. Guo, X. Luo, and H. Yu, "DyHDGE: Dynamic heterogeneous transaction graph embedding for safety-centric fraud detection in financial scenarios," *J. Saf. Sci. Resil.*, vol. 5, no. 4, pp. 486–497, Dec. 2024, doi: 10.1016/j.jnlssr.2024.05.005.
[25] K. Li *et al.*, "SEFraud: Graph-based Self-Explainable Fraud Detection via Interpretative Mask Learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug. 2024, pp. 5329–5338. doi: 10.1145/3637528.3671534.
[26] Y. Tang and Y. Liang, "Credit card fraud detection based on federated graph learning," *Expert Syst. Appl.*, vol. 256, p. 124979, Dec. 2024, doi: 10.1016/j.eswa.2024.124979.
[27] P. Grover *et al.*, "Fraud Dataset Benchmark and Applications," *ArXiv*. Sep. 22, 2023. [Online]. Available: http://arxiv.org/abs/2208.14417
[28] O. Ndama, S. Ndama, I. Bensassi, and E. M. En-Naimi, "A novel BERT-long short-term memory hybrid model for effective credit card fraud detection," *IAES Int. J. Artif. Intell.*, vol. 15, no. 1, p. 788, Feb. 2026, doi: 10.11591/ijai.v15.i1.pp788-797.
[29] A. Cherif, H. Ammar, M. Kalkatawi, S. Alshehri, and A. Imine, "Encoder–decoder graph neural network for credit card fraud detection," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 3, p. 102003, Mar. 2024, doi: 10.1016/j.jksuci.2024.102003.
[30] K. Shenoy, "Credit Card Transactions Fraud Detection Dataset," *Kaggle.com*, 2019. https://www.kaggle.com/datasets/kartik2112/fraud-detection