



A Segmentation-Aware Isolation Forest Framework for Fraud Investigation Prioritization in QRIS Transactions

Ahmad Rijaluddin Nisbil Kamal^{1*}, Valeriana Lukitosari²

^{1,2}Interdisciplinary School of Management and Technology,
Institut Teknologi Sepuluh Nopember, Indonesia

E-Mail: ¹rijal2602@gmail.com, ²valerianalukitosari@gmail.com

Received May 22th 2026; Revised Jun 08th 2026; Accepted Jun 17th 2026; Available Online Jul 18th 2026

Corresponding Author: Ahmad Rijaluddin Nisbil Kamal

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Fraud detection in Quick Response Code Indonesian Standard (QRIS) transaction environments is challenging because fraud labels are often limited, delayed, or incomplete, making it difficult to apply supervised learning effectively. In addition, high transaction volumes create operational constraints for investigation teams, requiring an efficient mechanism to prioritize high-risk merchants for review. To address this problem, this study proposes a segmentation-aware Isolation Forest framework for fraud investigation prioritization in QRIS transactions. Transaction-level data from a national bank in Indonesia were transformed into merchant-level behavioral representations using 38 engineered features capturing transaction intensity, temporal patterns, nominal concentration, burst behavior, and transactional irregularities. Anomaly scores were then used as ranking signals for prioritizing selective investigation. Model performance was evaluated using proxy-based ranking metrics, capacity simulation investigations, selective alert optimization, transaction-level interpretability, and expert validation. The results show that only 1.03% of merchants were classified as very high anomalies, while the model achieved strong ranking performance with an Area Under the Curve (AUC) of 0.8794, Precision@Top 5% of 0.3306, and Lift of 6.5942. Selective alerting reduced investigation workload by up to 80.24%, with higher precision and lower FAR, although at the cost of reduced recall. Expert validation further indicated that 94.12% of prioritized merchants were operationally relevant.

Keywords: *Fraud Investigation Prioritization, Isolation Forest, Merchant-Level Behavioral Profiling, QRIS Transactions, Segmentation-Aware Anomaly Detection.*

1. INTRODUCTION

The rapid expansion of digital payment systems has significantly transformed Indonesia's financial services landscape, with the Quick Response Code Indonesian Standard (QRIS) becoming one of the country's most important payment innovations since its introduction by Bank Indonesia in 2019 [1]. As the national payment standard in Indonesia, QRIS has accelerated transaction efficiency and financial inclusion in both the formal and informal sectors. By the second quarter of 2025, QRIS adoption had reached 39.3 million merchants—93.16% of which were Micro, Small, And Medium Enterprises (MSMEs)—serving 57 million users and processing 6.05 billion transactions with a total transaction value of IDR 579 trillion [2].

While this large-scale adoption creates substantial economic opportunities, it also introduces urgent challenges for fraud monitoring in high-volume transaction environments. In QRIS transactions, fraudulent activity does not always manifest as explicit rule violations, but often manifests as subtle deviations in transaction patterns, such as repeated transaction amounts, transactions outside of merchant operating hours, higher-than-average amounts for the merchant, spike transactions, and so on. These patterns are often adaptive and difficult to detect using conventional rules-based approaches, making fraud monitoring increasingly dependent on the ability to identify behavioral deviations at the merchant level.

This situation is reflected in an internal report from a national bank in Indonesia. According to their internal report, between August 2024 and August 2025, 682 fraud cases occurred by their merchants on one of their QRIS channels, referred to in this study as Channel A. These incidents resulted in losses of approximately IDR 800 million. As illustrated in Figure 1, fraud incidents increased significantly in early 2025, indicating an evolving fraud pattern that requires more adaptive and scalable monitoring mechanisms.

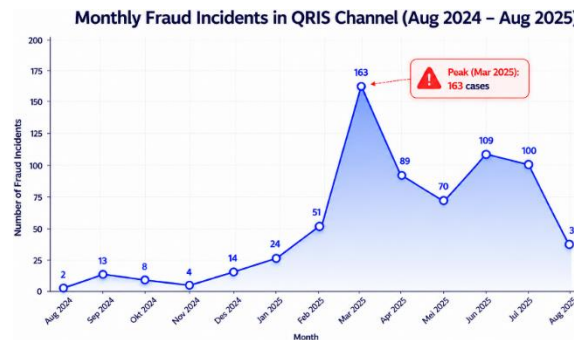


Figure 1. Monthly Fraud Incidents in QRIS Channel A (Aug 2024 – Aug 2025)

Despite these risks, fraud detection practices in many financial institutions remain largely reactive and rely heavily on predefined rules and manual reviews. While such approaches may detect known fraud patterns, they often fail to identify emerging behavioral anomalies in large-scale transaction environments [3], [4]. As transaction volumes continue to grow, this limitation creates an operational bottleneck because investigative teams cannot realistically review all suspicious transactions or merchants.

Recent studies further highlight that financial fraud detection is inherently challenged by extreme class imbalance, evolving fraud strategies, and the need for interpretable analytical systems capable of operating in highly dynamic financial environments [5].

In the real world, the number of fraud cases usually represents only a small fraction of total transactions. For example, prior work reported that only 492 out of 283,807 transactions were fraudulent in a real-world financial dataset [6]. Fraud labels are also often delayed, incomplete, or noisy due to verification costs, reporting delays, and uncertainty in confirmation processes [7], [8]. Under these conditions, traditional supervised learning approaches often struggle because they require high-quality labeled data [9], [10], while static rule-based systems may not adapt quickly to evolving fraud behavior [11]. Therefore, fraud detection in a payment environment with high transaction volumes should be treated not only as a classification problem, but also as a prioritization problem.

Several previous studies have demonstrated the relevance of anomaly detection and machine learning for financial fraud monitoring. Isolation Forest, introduced by Liu et al. (2008), is an efficient unsupervised anomaly detection method that isolates anomalies through random partitioning [12]. Ounacer et al. (2018) compared several unsupervised anomaly detection methods, including Local Outlier Factor, One-Class Support Vector Machine (SVM), K-Means, and Isolation Forest, on credit card transaction data and reported that Isolation Forest achieved strong detection performance with an Area Under the Curve (AUC) of 91% [13]. Recent studies also support the applicability of Isolation Forest in high-volume and dynamic data environments, including real-time online transaction fraud detection in digital banking and streaming data anomaly detection [14], [15]. Vanini et al. (2023) examined online payment fraud by linking anomaly detection with risk management, showing that fraud detection outputs should support risk-based decision-making rather than remain as isolated analytical results [11]. In a large-scale e-commerce setting, Nanduri et al. (2020) demonstrated that combining machine learning and optimization can support fraud reduction and improve operational transaction decision processes [16].

Although these studies show the value of anomaly detection and machine learning in fraud monitoring, their contexts differ from the problem addressed in this study. Prior works mainly focus on credit card transactions, online payment fraud, e-commerce systems, or general digital banking environments [6], [11], [13], [14], [15], [16]. These settings do not fully reflect the behavioral characteristics of QRIS merchant transactions, where risk may emerge from merchant-level activity patterns across heterogeneous transaction volumes, business profiles, and operating intensities. In addition, existing studies generally emphasize detection performance, risk modeling, model architecture, or transaction-level decision-making, while paying limited attention to how anomaly scores can be operationalized as a ranking mechanism for selective merchant investigation.

In QRIS transactions, merchants exhibit significant heterogeneity, especially in terms of transaction volume, activity intensity, and transaction characteristics. A single global anomaly model may produce biased scores because merchants with different transaction volumes and behavioral baselines are evaluated within the same population. This condition motivates the use of segmentation-aware modeling so that anomaly scores reflect deviations relative to comparable merchant groups. In this context, unsupervised anomaly detection is quite relevant to this problem. This is because unsupervised anomaly detection can generate relative risk signals without relying on a complete fraud label. However, these signals should be interpreted as indicators for further review, not as definitive proof that a merchant has committed fraud.

Motivated by these gaps, this study applies a segmentation-aware Isolation Forest model to merchant-level QRIS transaction data and evaluates anomaly scores as ranking signals for selective investigation rather

than direct fraud classification. Accordingly, this study aims to evaluate the effectiveness of unsupervised anomaly detection as a ranking mechanism for identifying QRIS merchants that require further review under limited fraud labels, and to examine how anomaly-based ranking can improve investigation prioritization and reduce operational alert workloads.

This research differs from previous studies by shifting the focus from transaction-level fraud detection, model performance optimization, or general online payment fraud analysis toward merchant-level investigation prioritization in QRIS transaction systems. The novelty of this study lies in operationalizing unsupervised anomaly detection as a segmentation-aware ranking framework that considers heterogeneous merchant behavior, proxy-based evaluation, selective alerting, and investigation capacity constraints. This study demonstrates how anomaly scores can be used not as final fraud labels but as practical prioritization signals to support fraud-risk monitoring and more efficient investigation workflows in QRIS environments. The remainder of this paper is organized as follows. Section II presents the methodology, Section III presents the empirical results and discusses the findings, and Section IV concludes the study.

2. Research Methodology

This study adopts a quantitative case study approach to examine how unsupervised anomaly detection can support prioritization of QRIS fraud investigations. As illustrated in Figure 2, the research workflow consists of seven stages: research context and data preparation, feature engineering, segmented Isolation Forest modeling, proxy-based prioritization evaluation, investigation workload optimization, behavioral interpretation and transaction drilldown, and expert validation.

The first stage defines the observation scope, merchant category coverage, unit of analysis, and merchant selection strategy. QRIS transaction logs are limited to selected MCC categories and observation periods. Because the transaction population is large, the study applies entity-based merchant selection using predefined criteria and transaction-volume strata to represent different merchant activity levels.

The second stage transforms transaction logs into merchant-level behavioral features through preprocessing, feature selection, and feature engineering. Preprocessing includes data cleaning, standardization, timestamp validation, numerical conversion, and identifier construction to support aggregation and drilldown analysis. Feature selection removes low-information and highly correlated variables, while feature engineering constructs behavioral indicators related to transaction intensity, amount patterns, temporal behavior, burst activity, recent shifts, and peer deviations.

The third stage applies segmented Isolation Forest modeling by grouping merchants according to transaction-volume segments to reduce heterogeneity bias. The model produces anomaly scores and percentile-based anomaly levels for each merchant. The fourth stage evaluates anomaly scores as prioritization signals using proxy behavioral risk indicators and ranking-oriented metrics, since verified fraud labels are unavailable.

The fifth stage evaluates investigation capacity and alert workload trade-offs through review-threshold simulations. The sixth stage provides behavioral interpretation and transaction-level drilldown by translating anomaly drivers into business-readable explanations and supporting transaction samples. Finally, expert validation assesses whether prioritized anomalies are operationally meaningful within a limited-label fraud monitoring environment.

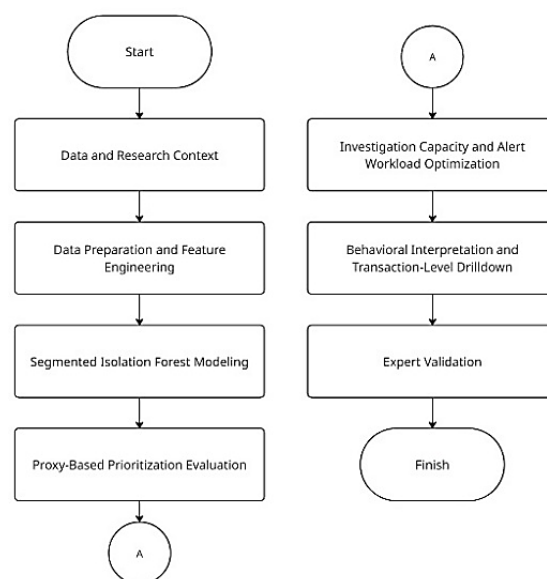


Figure 2. Conceptual Framework for Operationalizing Unsupervised Anomaly Detection for Fraud Investigation Prioritization

2.1. Data and Research Context

This study employs a quantitative case-study design using QRIS transaction logs from a national bank in Indonesia. The analysis focuses on merchant-level fraud investigation prioritization by identifying unusual behavioral patterns in QRIS transactions. Accordingly, anomaly scores are used as ranking signals for further review rather than direct fraud classification outputs.

The historical fraud context refers to incidents observed in QRIS Channel A from August 2024 to August 2025, where weaknesses in QRIS self-registration and transaction controls enabled abuse by certain merchants, including fictitious or inappropriate QRIS activity. These incidents resulted in customer losses and exposed the bank to operational, legal, and reputational risks. The bank subsequently strengthened its controls and related business processes.

The August 2024–August 2025 incidents serve as the operational motivation and contextual basis for the study, not as the empirical modeling dataset. The historical fraud patterns informed the 2026 analysis design, including the merchant-level unit of analysis, selected merchant categories, behavioral anomaly indicators, and the use of anomaly scores for investigation prioritization. However, historical fraud records were not used as supervised labels or direct training inputs for the Isolation Forest model.

The empirical dataset consists of QRIS transaction logs from January 26 to February 26, 2026, containing attributes such as transaction timestamp, amount, type, merchant identifier, merchant category code, and processing code. This post-improvement operational period was selected to evaluate whether merchant-level anomaly detection can support ongoing fraud-risk monitoring after the earlier Channel A incidents. The study does not estimate actual fraud prevalence because confirmed fraud labels may still be incomplete or delayed. Instead, it evaluates whether anomaly scores can identify merchants whose behavior deviates from comparable peers and should be prioritized for further review.

The study focuses on merchants under Grocery/Supermarkets (MCC 5411) and Miscellaneous Food Stores (MCC 5499), representing the daily necessities segment. These categories were selected because their relatively routine transaction behavior provides a clearer basis for detecting anomalies while remaining operationally relevant to the prior fraud-monitoring context. Since the unit of analysis is the merchant, transaction-level logs are transformed into merchant-level behavioral representations. The analysis is conducted offline, and anomaly scores are used to support investigation prioritization and fraud-risk monitoring rather than automatically classify merchants as fraudulent.

2.2. Data Preparation and Feature Engineering

The data preparation stage transforms raw QRIS transaction logs into merchant-level behavioral representations. The process begins with extracting transaction records from the operational database by applying the defined observation period and retaining selected MCC categories relevant to the study scope.

Because the raw transaction population is large, this study applies an entity-based merchant selection approach rather than random transaction sampling. Active merchants are selected based on predefined research criteria, and transaction-volume strata are used to ensure that merchants with different activity levels are represented in the analysis. All transaction records associated with the selected merchants are retained to preserve complete behavioral histories for feature construction and transaction-level drilldown.

The extracted data then undergo preprocessing procedures, including standardization of missing values and textual fields, timestamp validation, numerical conversion, and duplicate removal. A unique merchant identifier is constructed and used consistently across feature engineering, anomaly detection, transaction-level drilldown, and expert validation.

Transaction timestamps were converted into datetime format to derive temporal indicators, including transaction date, hour, day of week, weekend flag, and night-transaction flag. Transaction amounts were converted into numerical format, and a capped transaction amount variable was created by limiting extreme values at the 99.9th percentile. The original transaction amount was retained, while the capped amount was used to stabilize statistical feature construction and reduce excessive sensitivity to isolated extreme outliers.

To improve feature quality and reduce redundancy, a hybrid statistical and domain-screening approach was applied prior to feature engineering. Filter-based feature selection methods were used to remove features with near-zero variance and highly correlated feature pairs ($|r| > 0.90$) to reduce redundant information and multicollinearity [17]. Non-behavioral identifiers were also excluded to avoid artificial ranking bias.

Feature engineering was then performed to transform the selected variables into merchant-level behavioral representations. The engineered features captured transaction intensity, amount characteristics, temporal activity, nominal concentration, burst behavior, recent behavioral shifts, and deviations from MCC peer baselines. The final representation consisted of 38 numerical behavioral features.

2.3. Segmented Isolation Forest Modeling

The anomaly detection model uses Isolation Forest, where each merchant is represented as a numerical feature vector in a multidimensional behavioral space. The algorithm is selected because it identifies observations that are easier to isolate from the majority of normal observations. Unlike distance-based or

density-based approaches, Isolation Forest isolates anomalies through recursive random partitioning without explicit distance calculations. Anomalous observations tend to produce shorter isolation path lengths because they are behaviorally more distinct from normal populations [12], [18].

To account for transaction-volume heterogeneity, merchants were segmented based on transaction counts in the last five days of the observation window: High ($\geq 1,000$), Medium (200–999), Low (20–199), And Below_20 (< 20). For Each Segment, The Model Was Trained Using High_Confidence Merchants, Defined As Merchants With At Least 50 Transactions And 7 Active Days; Merchants Below These Criteria Were Still Scored But Marked As low_data. If A Segment Had Fewer Than 50 high-confidence merchants, the global high-confidence pool was used as a fallback, while segments with fewer than 20 training merchants were treated as unreliable for anomaly-level interpretation.

Numerical features were standardized before training, and Isolation Forest was configured with 300 estimators, contamination="auto", random_state=42, and parallel processing. The model output from score_samples was converted into anomaly scores by multiplying the normality score by -1, so higher values indicate stronger deviations. Final anomaly levels were assigned post-hoc using segment-specific percentile thresholds: Low ($< P95$), Medium ($P95-P97$), High ($P97-P99$), and Very_High ($\geq P99$).

Although more advanced variants, such as Extended Isolation Forest and Deep Isolation Forest have been proposed, they may introduce higher complexity and lower interpretability [18], [19]. Therefore, conventional iForest is considered more suitable for the operational setting of this study.

2.4. Proxy-Based Prioritization Evaluation

Because verified fraud labels are unavailable, model performance is evaluated using a proxy-based prioritization framework. Rather than evaluating anomaly detection as a direct fraud classifier, this stage assesses whether higher anomaly scores are associated with stronger behavioral risk indicators. This approach is aligned with foundational outlier analysis principles, which emphasize the importance of domain knowledge and context-specific behavioral features when explicit fraud labels are limited or unavailable [20].

A composite proxy risk score was constructed from suspicious behavioral indicators, including high transaction values, temporal irregularities, bursts, and repeated transaction amounts. Merchants in the top 5% of proxy scores were treated as proxy-positive, reflecting a realistic assumption about investigation capacity. Potential circularity was mitigated by threshold-independent metrics, including Spearman correlation and AUC.

Prioritization quality was assessed using Spearman correlation, Precision@K, Lift@K, recall, FAR, and AUC. These metrics evaluate whether anomaly ranking captures stronger proxy-risk merchants earlier than random review, with AUC interpreted cautiously because the benchmark is proxy-based rather than confirmed fraud labels. This evaluation strategy aligns with the study's objective: to assess anomaly scores as investigation-prioritization signals under limited-label conditions.

2.5. Investigation Capacity and Alert Workload Optimization

Investigation capacity was evaluated using simulated review thresholds. Merchants were ranked by anomaly score, and top-ranked subsets at 1%, 3%, 5%, and 10% were analyzed to assess the trade-off between workload and proxy-risk coverage.

Selective alerting was also evaluated using three scenarios: a global top-percentile threshold, segment-specific thresholds, and a stricter high-priority alert layer. These scenarios were compared using alert volume, precision, recall, lift, and FAR to assess whether anomaly detection can reduce unnecessary workload while preserving prioritization quality.

2.6. Behavioral Interpretation and Transaction-Level Drilldown

Post-hoc behavioral interpretation was conducted after model scoring. This layer does not change the anomaly score, but translates dominant deviations into business-readable reason codes. Dominant behavioral dimensions were identified using deviations from segment-level baselines, including high-value behavior, nominal concentration, temporal concentration, burst activity, recent shifts, and MCC benchmark deviations.

Transaction-level drilldown was conducted for Very_High anomaly merchants with High_Confidence to connect merchant-level rankings with concrete transaction evidence. Candidate transactions were identified using driver-aware rules, such as extreme transaction values, repeated amounts, night-time activity, short gaps, and peak windows. This step produces transaction samples that help explain why a merchant was prioritized and provides supporting evidence for expert review.

2.7. Expert Validation

To assess whether prioritized anomalies are operationally meaningful, expert validation is conducted using a structured review pack. The validation process consists of three components. First, the model output provides ranked merchant-level anomaly results, including scores, segments, anomaly levels, and behavioral explanations. Second, the transaction sample provides transaction-level drilldown evidence for Very_High

anomaly merchants with High_Confidence. Third, the SME review records subject matter expert judgments on whether each prioritized merchant is suspicious, requires monitoring, or appears normal. This validation does not replace confirmed fraud labels, but serves as an operational relevance check in a limited-label environment.

3. RESULTS AND DISCUSSION

3.1. Data Characteristics and Feature Representation

This study analyzes QRIS transaction logs from one of Indonesia's national banks covering the period from January 26 to February 26, 2026. Following preprocessing, merchant filtering, and entity-based merchant selection, the final modeling dataset consisted of 4,952 merchants and 10,052,810 transactions, focusing on merchants within the daily necessities segment, namely MCC 5411 and MCC 5499. This segment was selected because it provides relatively routine transaction behavior while also representing an operational area where suspicious activities had previously been observed to be more concentrated.

The empirical analysis reveals three key characteristics of QRIS merchant behavior. First, transaction values are strongly asymmetric, with most transactions concentrated at lower nominal values and a small proportion extending to much higher amounts. Second, transaction activity shows temporal dependence, with 74.8% occurring on weekdays and 25.2% on weekends. Third, cross-merchant heterogeneity is substantial: each merchant processed approximately 2,030 transactions on average, but transaction intensity varied considerably and did not scale linearly with transaction values.

These characteristics indicate that merchant prioritization requires multivariate behavioral representations that jointly capture transaction value patterns, temporal activity, transaction intensity, nominal concentration, and peer-baseline deviations. They also support the use of segmentation-aware anomaly detection, as merchants with different transaction volumes may naturally exhibit distinct behavioral baselines.

Transaction-level observations were therefore aggregated into merchant-level behavioral representations. From 41 initial variables, statistical filtering and domain validation reduced the feature space to 17 core numerical variables, which were expanded into 38 behavioral features. Table 1 summarizes the engineered feature groups used to represent merchant behavior.

Table 1. Feature Engineering

Feature Group	Features	Brief Description
Volume and Activity	trx_cnt, active_days, trx_per_day	Capture merchant transaction intensity and activity consistency.
Transaction Amount Profile	avg_trx_amount, median_trx_amount, std_trx_amount, p99_trx_amount	Describe typical, variable, and extreme transaction amounts.
Refund and Tip Behavior	refund_ratio, avg_tip_amount, pct_trx_with_tip	Capture refund and tip-related transaction patterns.
Temporal Behavior	pct_trx_night, pct_trx_weekend, entropy_trx_hour	Describe transaction timing across night, weekend, and hourly patterns.
High-Value Behavior	pct_high_value_trx	Measure the share of high-value transactions.
Nominal Concentration and Repetition	top1_amount_share, top3_amount_share, entropy_trx_amount, dominant_amount_gap, same_amount_streak, repeated_amount_ratio	Capture repeated amounts and nominal-value concentration.
Burst and Interarrival Pattern	peak_hour_ratio, median_interarrival_sec, p05_interarrival_sec, burst_short_gap_ratio, max_trx_per_30min	Represent transaction density, speed, and short-window bursts.
Relative Amount Extremes	p95_to_median_ratio, p99_to_median_ratio, max_to_median_ratio, min_to_median_ratio	Measure extreme-value deviations from typical transaction amounts.
Daily Instability and Spikes	spike_ratio, cv_daily_trx	Capture daily transaction volatility and spikes.
Composite Temporal Risk	weekend_night_interaction_ratio	Measure combined weekend and night-time activity concentration.
Recent Behavioral Shift	last_7d_trx_vs_prev_ratio, last_7d_amt_vs_prev_ratio, last_7d_night_ratio_delta, last_7d_weekend_ratio_delta, last_7d_high_value_ratio_delta	Capture recent changes relative to prior merchant behavior.
MCC Benchmark Deviation	avg_amt_vs_mcc	Measure deviation from peer merchants within the same MCC.

Table 1 shows that the feature design captures multiple dimensions of merchant behavior, including transaction intensity, amount variation, temporal activity, nominal concentration, burst patterns, and deviations from MCC peer baselines. These features provide the behavioral foundation for anomaly detection and support the use of anomaly scores as prioritization signals rather than direct fraud labels.

3.2. Segmentation-Aware Anomaly Detection Results

Given the high heterogeneity across merchants, anomaly detection was performed separately across four transaction-volume segments, as defined in the methodology: Low, Medium, High, and Below_20. Table 2 presents the distribution of merchants across these segments before the high-confidence filtering used in model evaluation.

Table 2. Merchant Segment Distribution

Segment	Number of Merchants
Low	2,998
Medium	1,552
High	64
Below_20	338
Total	4,952

Table 2 shows that the merchant population is dominated by Low-volume merchants, with 2,998 merchants, followed by 1,552 merchants in the Medium segment. Meanwhile, only 64 merchants belong to the High segment, and 338 merchants fall into the Below_20 segment. This distribution confirms substantial variation in merchant activity levels and supports the need for segmentation-aware anomaly detection to reduce bias from applying a single global comparison across heterogeneous merchants.

Based on the segment-specific percentile thresholds defined in the methodology, merchants were categorized into Low, Medium, High, and Very_High anomaly levels. The resulting anomaly distribution is presented in Table 3.

Table 3. Anomaly Distribution

Anomaly Level	Merchants	Percentage
LOW	4,698	94.87%
MEDIUM	101	2.04%
HIGH	102	2.06%
VERY_HIGH	51	1.03%

Table 3 shows that anomalous behavior is concentrated within a relatively small subset of merchants. Most merchants (94.87%) were classified as Low anomaly, while only 1.03% were categorized as Very_High anomaly. This indicates that the model produces selective prioritization signals and can narrow investigation targets to merchants with stronger behavioral deviations. However, these anomaly levels should be interpreted as signals for investigation prioritization rather than confirmed fraud labels.

The concentration of higher anomaly levels within a small merchant subset supports the use of anomaly scores as a filtering mechanism for investigation prioritization. This result is consistent with the principle of Isolation Forest, where anomalies are expected to represent a small fraction of observations [12], and with prior findings showing that Isolation Forest can identify anomalous financial transactions without labeled fraud data [13]. In QRIS fraud monitoring, this selectivity has operational value because investigators can focus on a limited set of higher-priority merchants instead of reviewing broad alert populations generated by static rules.

3.3. Prioritization Performance Evaluation

Since verified fraud labels were unavailable, model performance was evaluated using proxy-based risk indicators. Merchants in the top 5% of proxy scores were classified as proxy-positive merchants, resulting in 248 proxy-positive merchants among 4,946 high-confidence merchants. The number of high-confidence merchants differs slightly from the full modeling population of 4,952 because the evaluation stage only includes merchants with sufficient behavioral evidence for a reliable proxy-based assessment.

Table 4 presents the main evaluation results used to assess whether anomaly scores can function as prioritization signals. The metrics are ranking-oriented rather than classification-oriented because the objective of this study is to prioritize merchants for review, not to classify merchants directly as fraudulent or non-fraudulent.

Table 4 shows that anomaly scores align well with proxy-risk indicators. Precision@Top 5% and Recall@Top 5% both reached 0.3306, indicating that the top-ranked merchants captured approximately one-third of all proxy-positive merchants. The model also achieved a Lift score of 6.5942 and an AUC of 0.8794,

demonstrating strong ranking capability compared with random review, while maintaining a relatively low False Alarm Rate of 0.0353.

Table 4. Main Evaluation Results

Metric	Value
Spearman (score vs proxy score)	0.4897
Spearman (score vs proxy label)	0.2868
Precision@Top 5%	0.3306
Recall@Top 5%	0.3306
FAR	0.0353
AUC	0.8794
Lift	6.5942

These results indicate that anomaly scores are effective for prioritizing merchants with stronger behavioral risk signals, although they should still be interpreted as review priorities rather than direct fraud classifications. To further assess ranking reliability across heterogeneous merchant groups, segment-level performance was evaluated using high-confidence merchants only. Table 5 presents the performance of the anomaly ranking across merchant-volume segments.

Table 5. Segment-Level Performance

Segment	N	Precision	Recall	FAR	AUC
LOW	2,994	0.3111	0.2276	0.0216	0.9032
MEDIUM	1,552	0.3462	0.2596	0.0352	0.8352
HIGH	64	0.4000	0.3333	0.0517	0.8966
BELOW_20	336	0.0000	0.0000	0.0125	0.8881

Table 5 shows that all merchant segments maintained strong AUC performance above 0.83, indicating that anomaly scores preserved useful ranking information across heterogeneous merchant-volume groups. The Low and High segments achieved AUC values above 0.89, while the Medium segment reached 0.8352.

The segment-level results also show relatively stable ranking performance across merchant groups. The Below_20 segment presents a special case where precision and recall were both zero under the selected segmented threshold, while the AUC remained high at 0.8881. This should not be interpreted as a failure of the anomaly ranking signal. In the segmented thresholding procedure, precision, recall, and FAR were derived from binary alert decisions using segment-specific score cutoffs, whereas AUC was calculated directly from continuous anomaly scores and was therefore threshold-independent.

The Below_20 segment applied the most conservative cutoff at the 99th percentile because low-activity merchants had fewer behavioral observations and were more prone to unstable alerts. As a result, no proxy-positive merchants were included in the alert set, producing zero precision and recall. Nevertheless, the high AUC indicates that the anomaly scores still preserved useful ranking information within this segment, although threshold-based results for low-activity merchants should be interpreted cautiously.

3.4. Prioritization Optimization Under Limited Investigation Capacity

To evaluate practical scalability, investigation capacity simulations were conducted by varying review thresholds. Table 6 presents the number of merchants reviewed, the number of proxy-positive merchants captured, the capture rate, precision, and lift under four review-capacity scenarios: top 1%, top 3%, top 5%, and top 10%.

Table 6. Investigation Capacity Simulation

Review Capacity	Merchants Reviewed	Proxy Captured	Capture Rate	Precision	Lift
Top 1%	50	16	0.0645	0.3200	6.3819
Top 3%	149	48	0.1935	0.3221	6.4248
Top 5%	248	82	0.3306	0.3306	6.5942
Top 10%	495	130	0.5242	0.2626	5.2377

Table 6 shows a clear trade-off between investigation workload and risk coverage. Reviewing only the top 1% of merchants captured 16 proxy-positive merchants, while expanding the review threshold to 10% captured 130 proxy-positive merchants or 52.42% of all proxy-positive merchants. Smaller review thresholds produced higher lift values, whereas larger thresholds increased overall risk coverage.

This trade-off is further illustrated in Figure 3. The cumulative gains curve compares the proportion of proxy-risk merchants captured by anomaly-based ranking against a random-review baseline.

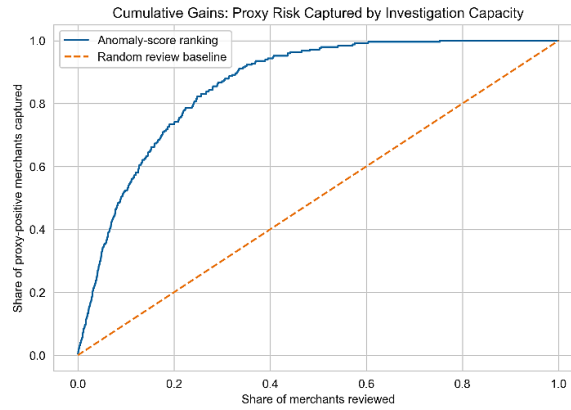


Figure 3. Cumulative Gains: Proxy Risk Captured by Investigation Capacity

Figure 3 shows that the anomaly-based ranking captures proxy-positive merchants substantially faster than random review. The steep early rise of the curve indicates that behavioral risk is concentrated near the top anomaly rankings, allowing investigators to identify more relevant merchants within limited review capacity. These findings demonstrate that anomaly-based ranking can support flexible investigation strategies by balancing operational workload and risk coverage according to institutional capacity.

3.5. Operational Efficiency Through Selective Alerting

Beyond ranking quality, anomaly detection must also reduce unnecessary operational workload. As shown in Table 7, applying a global threshold generated 248 alerts, requiring investigators to review approximately 5.01% of all merchants. Segment-specific thresholds reduced alerts to 177 merchants, representing a 28.63% reduction in workload, while maintaining relatively stable precision at 0.3220. A stricter high-priority alert layer further reduced alerts to only 49 merchants, representing an 80.24% reduction in investigation workload, while simultaneously achieving the highest precision at 0.3469 and the lowest FAR at 0.0068. These findings highlight that the value of anomaly detection lies not in increasing alert volume, but in improving the precision of investigation targeting while significantly reducing operational workload.

Table 7. Operational Workload Comparison

Policy	Alerts	Precision	Recall	FAR	Lift
Global	248	0.3306	0.3306	0.0353	6.5942
Segmented	177	0.3220	0.2298	0.0255	6.4225
High Priority	49	0.3469	0.0685	0.0068	6.9192

3.6. Behavioral Interpretation of Prioritized Anomalies

For anomaly prioritization to be operationally useful, investigators must understand why specific merchants were flagged. Therefore, this study incorporates transaction-level drilldown to provide observable evidence behind prioritized anomaly scores. Table 8 presents sample transaction patterns from prioritized merchants after masking merchant identifiers.

Table 8. Sample Transaction Pattern

Merchant ID (Masked)	Transaction Amount (IDR)	Transaction Hour	Key Indicator
10458***	1,000,000	05	Extreme value + night transaction
10458***	1,000,000	07	Extreme value deviation
10458***	700,000	09	Above normal merchant range
10458***	10,000	04	Repeated low-value transactions
10458***	5,000	05	Repeated transactions within a short interval
10458***	20,000	04	Recurrent early-morning transaction pattern

As shown in Table 8, the transaction-level drilldown provides additional evidence of how anomaly signals emerge at the operational level. Two recurring behavioral patterns can be observed. First, several transactions significantly exceeded the merchant's normal transaction range, including transactions of IDR 1,000,000 and IDR 700,000, indicating abnormal value deviations relative to historical merchant behavior. Second, several transactions occurred during early-morning hours, particularly around 04:00 and 05:00, indicating unusual temporal concentration. These findings show that top-ranked merchants were prioritized

due to a combination of value deviation and off-hour transaction activity, rather than solely due to large transaction values.

These findings show that top-ranked merchants were prioritized due to combinations of value deviation, nominal repetition, and recurring off-hour activity, rather than solely due to large transaction values. This supports the use of multivariate behavioral features in anomaly-based ranking. The interpretation layer also improves operational explainability. Instead of only producing anomaly scores, the framework provides observable evidence such as repeated nominal transactions, unusual timing, and abnormal transaction values, thereby helping bridge analytical outputs and investigator decision-making.

3.7. Expert Validation of Prioritized Alerts

Finally, expert validation was conducted to assess whether prioritized merchants were operationally meaningful. The validation process involved three structured review components: anomaly ranking outputs from Model_Output, transaction-level evidence from Transaction_Sample, and final business review through SME_Review_Business. Table 9 summarizes the expert assessment results for Very_High anomaly merchants.

Table 9. Expert Validation Summary

Assessment	Merchants	Percentage
Suspicious	21	41.18%
Requires Monitoring	27	52.94%
Normal	3	5.88%

As shown in Table 9, experts reviewed all 51 Very_High anomaly merchants and classified 21 merchants, or 41.18%, as suspicious; 27 merchants, or 52.94%, as requiring monitoring; and only 3 merchants, or 5.88%, as normal. Overall, 94.12% of prioritized merchants were considered operationally relevant by experts, while the small normal category primarily involved internal testing merchants whose unusual behavior was expected. This result indicates that anomaly-based prioritization aligns with real-world investigative judgment and can serve as a decision-support layer to help investigators focus on merchants that require earlier review.

4. CONCLUSION

This study demonstrates that unsupervised anomaly detection can support prioritizing QRIS fraud investigations when confirmed fraud labels are limited, delayed, or incomplete. By applying a segmentation-aware Isolation Forest framework, transaction-level QRIS data were transformed into merchant-level behavioral representations, enabling merchants to be ranked based on multidimensional behavioral deviations rather than isolated transaction rules.

The strength of this study lies in repositioning anomaly detection as an operational prioritization mechanism rather than a direct fraud classification model. The proposed framework combines merchant-level feature engineering, transaction-volume segmentation, proxy-based evaluation, selective alerting, transaction-level drilldown, and expert validation. The results show that anomaly scores can function as ranking signals to identify a small subset of merchants requiring further review, support more focused investigation, and reduce alert workload through selective prioritization.

This study has several limitations. The analysis is limited to active merchants in MCC 5411 and MCC 5499 within the daily necessities segment and is based on entity-based merchant selection with volume- and MCC-based stratification. Therefore, the findings should be interpreted within the selected merchant scope and may not fully generalize to other merchant categories or the entire QRIS transaction population. In addition, model evaluation relied on proxy risk indicators because verified fraud labels were unavailable during the modeling period. Therefore, anomaly scores should be interpreted as signals for investigation prioritization, not as confirmed fraud classifications.

Future research should incorporate verified fraud outcomes, broader MCC categories, longer observation periods, and multi-institution datasets to improve validation and generalizability. Further improvements may include real-time streaming, adaptive thresholds, graph-based analysis of merchant relationships, and investigator feedback loops to continuously refine risk indicators and prioritization strategies. Overall, the proposed framework provides an interpretable and scalable decision-support approach for QRIS fraud-risk monitoring by helping institutions prioritize merchants that require earlier review under constrained investigation capacity.

REFERENCES

- [1] Bank Indonesia: *Menavigasi Sistem Pembayaran Nasional di Era Digital*. Jakarta: Bank Indonesia, 2019.
- [2] Bank Indonesia, "QRIS Jelajah Indonesia 2025 Dorong Digitalisasi dengan Wisata Budaya," Indonesia, 2025. [Online]. Available: <https://www.bi.go.id/id/publikasi/ruang-media/news->

- release/Pages/sp_2717025.aspx. [Accessed: Oct. 10, 2025].
- [3] D. Kumar and S. Singh, "Analyzing the Impact of Machine Learning Algorithms on Risk Management and Fraud Detection in Financial Institution," *Int. J. Res. Publ. Rev.*, vol. 5, no. 5, pp. 1797–1804, 2024, doi: 10.55248/gengpi.5.0524.1135.
 - [4] J. de J. Rocha-Salazar, M. J. Segovia-Vargas, and M. del M. Camacho-Miñano, "Money laundering and terrorism financing detection using neural networks and an abnormality indicator," *Expert Syst. Appl.*, vol. 169, p. 114470, 2021, doi: 10.1016/j.eswa.2020.114470.
 - [5] A. A. Compagnino *et al.*, "An Introduction to Machine Learning Methods for Fraud Detection," *Appl. Sci.*, vol. 15, no. 21, pp. 1–32, 2025, doi: 10.3390/app152111787.
 - [6] M. Abdurrochman *et al.*, "Optimized Machine Learning Model for Credit Card Fraud Detection Using Smote-Tomek and Feature Engineering," *J. Appl. Informatics Comput.*, vol. 8, no. 2, pp. 580–588, 2024, doi: 10.30871/jaic.v8i2.8732.
 - [7] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system : A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, 2016, doi: 10.1016/j.jnca.2016.04.007.
 - [8] A. D. Pozzolo, G. Boracchi, O. Caelen, and C. Alippi, "Credit Card Fraud Detection : A Realistic Modeling and a Novel Learning Strategy," *IEEE Trans. NEURAL NETWORKS Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2017, doi: 10.1109/TNNLS.2017.2736643.
 - [9] M. R. Reddy, H. Q. Owaied, M. M. Abdulhasan, P. Karthik, M. Rajkumar, and P. U. Kumar, "Fraud Detection of Online transaction Using Machine Learning," *2024 Int. Conf. Augment. Reality, Intell. Syst. Ind. Autom. (ARIIA), Manipal, India, 2024*, pp. 1–7, 2024, doi: 10.1109/ARIIA63345.2024.11051851.
 - [10] M. A. Talukder *et al.*, "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," *J. Big Data*, vol. 11, no. 1, p. 33, 2024, doi: 10.1186/s40537-024-00886-w.
 - [11] P. Vanini, S. Rossi, E. Zvizdic, and T. Domenig, "Online payment fraud: from anomaly detection to risk management," *Financ. Innov.*, vol. 9, no. 1, p. 66, 2023, doi: 10.1186/s40854-023-00470-w.
 - [12] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," *2008 Eighth IEEE Int. Conf. Data Mining, Pisa, Italy*, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.
 - [13] S. Ounacer, H. A. El Bour, Y. Oubrahim, M. Y. Ghomari, and M. Azzouazi, "Using Isolation Forest in anomaly detection: The case of credit card transactions," *Period. Eng. Nat. Sci.*, vol. 6, no. 2, pp. 394–400, 2018, doi: 10.21533/pen.v6i2.533.
 - [14] H. Abbassi, S. E. L. Mendili, and Y. Gahi, "Digital banking fortification: a real-time isolation forest architecture for detecting online transaction fraud," *Eng. Res. Express*, vol. 6, no. 2, p. 25214, 2024, doi: 10.1088/2631-8695/ad4958.
 - [15] M. S. Kareem and L. A. N. Muhammed, "Anomaly Detection in Streaming Data using Isolation Forest," *2024 Seventh Int. Women Data Sci. Conf. Prince Sultan Univ. (WiDS PSU), Riyadh, Saudi Arab.*, pp. 223–228, 2024, doi: 10.1109/WiDS-PSU61003.2024.00052.
 - [16] J. Nanduri, Y. Jia, A. Oka, J. Beaver, and Y. Liu, "Microsoft Uses Machine Learning and Optimization to Reduce E-Commerce Fraud," *INFORMS J. Appl. Anal.*, vol. 50, no. 1, pp. 64–79, 2020, doi: 10.1287/inte.2019.1017.
 - [17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 1st ed. New York: Springer, 2013. doi: 10.1007/978-1-4614-6849-3.
 - [18] Y. Chabchoub, M. U. Togbe, A. Boly, and R. Chiky, "An In-Depth Study and Improvement of Isolation Forest," *IEEE Access*, vol. 10, pp. 10219–10237, 2022, doi: 10.1109/ACCESS.2022.3144425.
 - [19] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12591–12604, 2023, doi: 10.1109/TKDE.2023.3270293.
 - [20] C. C. Aggarwal, *Outlier Analysis*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-47578-3.