BulkShield: A Proactive Framework for Detecting Bulk Refund Booking in Ticketing Systems

Seungyoung Shin
Department of Data Science
Hanyang University
Seoul, Republic of Korea
seungyoungg@hanyang.ac.kr

Donggeun Oh
Department of Artificial Intelligence
Hanyang University
Seoul, Republic of Korea
dgoh1620@hanyang.ac.kr

Yongho Jeon
Department of Sales Planning
SR Corporation
Seoul, Republic of Korea
yh.j@srail.or.kr

Jinju Noh
Department of Sales Planning
SR Corporation
Seoul, Republic of Korea
jj.noh@srail.or.kr

Iltaeck Joo
Department of Data Science
Hanyang University
Seoul, Republic of Korea
itjoo@hanyang.ac.kr

Youngtae Noh*
Department of Data Science
Hanyang University
Seoul, Republic of Korea
youngtaenoh@hanyang.ac.kr

Abstract

Bulk Refund Booking (BRB) is a system-level demand distortion phenomenon in railway ticketing systems, where repeated strategic booking–refund behaviors create artificial seat shortages and revenue loss despite sufficient physical capacity. Unlike explicit rule violations, BRB emerges through institutionally permitted actions, making its risk boundaries difficult to define and limiting existing post-hoc threshold-based enforcement in supporting proactive intervention. To address these challenges, we conduct the first systematic empirical study of BRB using large-scale real-world railway ticketing logs and show that BRB risk emerges as a cumulative, sequential behavioral process rather than as isolated refund events. We propose *BulkShield*, an end-to-end proactive framework that combines unsupervised risk pseudo-labeling, booking-time sequence modeling, and LLM-assisted explanations grounded in model-identified behavioral evidence. Expert-in-the-loop evaluation shows strong alignment between learned risk boundaries and practitioner judgment ($\kappa = 0.8862$), validating the data-driven BRB risk structure. BulkShield achieves a Macro F1 score of 0.8909, outperforming strong sequential baselines, while practitioner feedback supports the usefulness of its natural-language explanations for operational intervention. These results highlight how behavior-aware and interpretable AI can uncover emerging demand distortion mechanisms and support proactive, evidence-based management in public transportation systems.

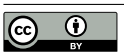
CCS Concepts

• **Information systems → Data mining**.

Keywords

Ticketing Systems; Fraud Detection; Large Language Models

*Corresponding author



KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818875

ACM Reference Format:
Seungyoung Shin, Donggeun Oh, Yongho Jeon, Jinju Noh, Iltaeck Joo, and Youngtae Noh. 2026. BulkShield: A Proactive Framework for Detecting Bulk Refund Booking in Ticketing Systems. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818875

KDD Availability Link: The source code for BulkShield is publicly available at https://github.com/imc-hanyang/BulkShield.

1 Introduction

In Korea's Super Rapid Train (SRT) system [37], which serves over 26 million passengers annually [40], flexible refund policies and mobile/web-based ticketing services [39] can enable **Bulk Refund Booking (BRB)**, where users strategically book tickets in bulk and later refund them at scale. As illustrated in Fig. 1, BRB temporarily holds seats through repeated booking–refund patterns, making observed demand appear higher than actual seat use and leading to artificial seat shortages and revenue loss despite sufficient physical capacity. The resulting inefficiency is substantial: refunded seats increased from 5.8 million in 2020 to 10.8 million in 2024 [5], 6.6% of refunded seats remained unsold, and artificial seat shortages affected nearly 20% of daily passenger demand [16], despite sufficient physical capacity [29].

Despite these growing impacts, existing post-hoc and rule-based detection approaches (*e.g.*, monthly refunds exceeding 1 million KRW, approximately USD 750, or refund rates above 90% [38]) remain inadequate for early BRB prevention because they rely on static indicators observed only after large refund behaviors have already accumulated [4]. These threshold-based criteria can also be avoided through minor behavioral adjustments, limiting their preventive effect in practice [9]. More fundamentally, they fail to capture how BRB risk emerges from cumulative and sequential booking–refund behaviors, leaving early-stage signals obscured. This motivates proactive detection approaches that model how BRB risk evolves over time rather than relying solely on post-hoc violations [6].

However, detecting BRB in public transportation systems remains challenging with existing behavioral risk detection approaches,

which are typically trained on instances labeled using well-defined risk criteria and produce coarse binary decisions [46] or risk scores [19]. Rather than constituting clear rule violations, BRB arises from the strategic use of institutionally permitted booking and refund procedures, placing it in a gray area where reliable labels are inherently difficult to define. Moreover, in public-sector transportation systems, detection outcomes must directly support passenger sanctions and policy interventions, requiring explanations that go beyond coarse risk scores to justify operational actions.

To address these challenges, we propose **BulkShield**, an end-to-end proactive and interpretable framework for defining, detecting, and explaining BRB risk in real-world railway operations. BulkShield derives unsupervised pseudo-labels to define data-driven risk boundaries, models booking–refund histories as temporal sequences for booking-time detection, and generates LLM-assisted explanations grounded in model-identified behavioral evidence. Through quantitative experiments and expert evaluation with transportation practitioners, we show that BulkShield achieves strong proactive detection performance, produces practitioner-aligned risk boundaries, and provides explanations that support operational decision-making.

- We present the first systematic empirical study of *Bulk Refund Booking* as a system-level demand distortion phenomenon in a real-world railway ticketing system, showing that BRB risk emerges through cumulative and sequential booking–refund behaviors rather than isolated post-hoc violations.
- We propose *BulkShield*, an end-to-end framework that addresses label scarcity, booking-time detection, and interpretability by integrating unsupervised data-driven risk labeling, sequence-based BRB detection, and evidence-grounded natural-language explanations.
- We validate BulkShield through quantitative experiments and expert evaluation with transportation practitioners, demonstrating strong proactive detection performance and practitioner-aligned risk boundaries and explanations beyond predictive accuracy alone.

2 Related Work

2.1 Fraud Detection

Most existing fraud detection studies formulate the task as a binary decision problem, assuming that fraud is a clearly identifiable event with a stable boundary [19, 27]—such as stolen card usage or account takeover—that enables labels to be directly used for detection modeling in domains like finance [46] and e-commerce [47]. Based on such boundary, prior studies have explored various detection strategies, including supervised models trained on labeled normal and fraudulent instances [14], unsupervised or weakly supervised methods that learn the distribution of normal behavior and consider deviations as potential fraud [22]. However, bulk refund booking lacks an equivalent, externally verifiable criterion for defining fraudulent behavior, which makes it difficult to directly apply existing fraud detection approaches built on fixed normal–risk boundaries. To bridge this gap, we construct data-driven pseudo-labels that define intermediate risk levels and validate their practical relevance through domain-expert assessment before using them for detection modeling.

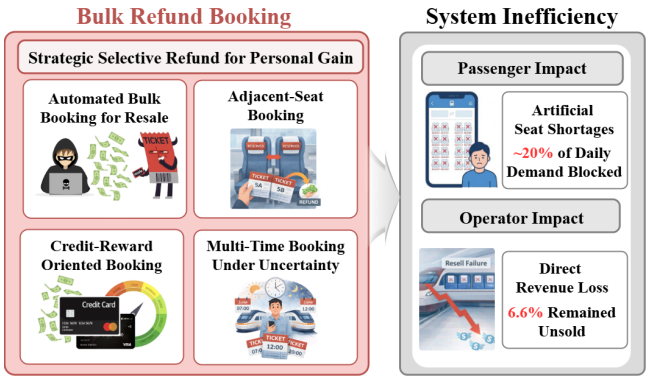


Figure 1: Overview of Bulk Refund Booking (BRB) as a system-level demand distortion phenomenon in railway ticketing systems. Strategic booking–refund behaviors, detailed in Sec. 3.1, temporarily withhold seats from passenger access and distort observed demand signals, leading to artificial seat shortages and operator revenue loss despite sufficient physical capacity.

2.2 Post-Hoc Explanation

To support the practical adoption of complex deep learning models, post-hoc explainability that interprets prediction processes has become imperative. As a result, post-hoc explanation methods have been extensively studied [11], including perturbation-based (e.g., LIME [30], SHAP [20]) and gradient-based (e.g., Integrated Gradients [41], Saliency Maps [34]) approaches which primarily yield quantitative feature attributions. However, their reliance on quantitative attributions often fails to provide sufficient semantic interpretability in practical settings, limiting their usefulness for practitioner decision-making [17, 25]. Recent studies address this by leveraging the semantic reasoning of LLMs [13, 44] to translate quantitative indicators—such as feature importance into semantic explanations, which have been shown to improve practitioner decision-making efficiency [15, 36]. Building on prior work, we advance post-hoc explanation for bulk refund booking detection by moving beyond feature-level attribution to sequence-level interpretation, which is essential for capturing temporally accumulated behaviors. Specifically, we identify prediction-critical events from the sequence model's inference process and incorporate them into attention-guided prompting, enabling LLMs to generate interpretable explanations grounded at the event-sequence level.

3 Preliminary

This section introduces the operational interpretation of BRB behavior and formalizes the proactive user-level detection task with explanation generation.

3.1 Bulk Refund Booking Patterns and Operational Interpretation

Beyond detecting whether a user is at BRB risk, effective operational handling requires interpreting what the accumulated booking–refund behavior represents. Through discussions with SR experts,

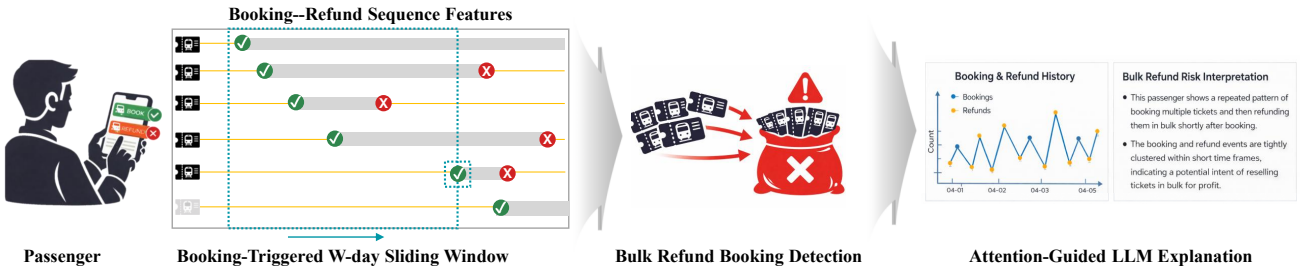


Figure 2: Overview of BulkShield. For each booking event, BulkShield extracts booking–refund sequence features from a (W)-day sliding window, predicts booking-time BRB risk, and generates expert-oriented explanations grounded in attention-guided evidence events. The window is updated as new bookings occur, enabling continuous booking-time risk assessment.

we identify two complementary interpretation needs for practitioner review: identifying the BRB-like risk pattern reflected in the observed evidence and contextualizing the behavior within the user's booking history to assess whether it is more consistent with strategic misuse or normal booking contexts.

We therefore define representative BRB pattern types observed in practice: (i) *Adjacent-Seat Booking with Selective Refund*, in which adjacent seats are reserved to secure personal space and selectively refunded shortly before departure [5]; (ii) *Automated Bulk Booking for Resale with Partial Utilization*, involving large-scale bookings that are later partially used or refunded, often in connection with resale attempts [12]; (iii) *Multi-Time Booking Under Uncertain-Schedule Booking*, in which multiple candidate travel times are reserved and most are later refunded [5]; and (iv) *Credit-Reward–Oriented Booking with Refund*, in which non-travel-driven bookings are made to accumulate card rewards or loyalty benefits and later refunded without actual travel [28]. However, BRB pattern identification alone is insufficient for operational interpretation, because similar booking–refund sequences may arise from different booking contexts.

To support booking-context interpretation, we further define representative purpose-driven contexts through expert consultation: (i) *Routine Commuting*, stable weekday travel between fixed origin–destination pairs; (ii) *Weekday Out-of-Region Stay with Weekend Return Travel*, Friday departures with Sunday or Monday returns, often work-related; (iii) *Leisure-Oriented Travel*, irregular weekend-centered trips across diverse routes; (iv) *Medical Travel*, periodic visits to hospital-adjacent stations, often involving weekend travel to metropolitan areas; (v) *Occasional Use*, sparse ad-hoc bookings under schedule uncertainty; and (vi) *Corporate Booking*, tickets reserved on behalf of multiple employees. These purpose-driven contexts are not used as risk labels, model inputs, or enforcement criteria; instead, they provide an interpretive reference frame for assessing whether observed booking–refund behaviors are more consistent with normal use or suggestive of strategic misuse, thereby supporting explanation and operational judgment.

3.2 Problem Formulation

We formulate proactive Bulk Refund Booking detection as a user-level binary classification problem triggered at each booking event. For each user i, given a booking event at time t, we construct an event sequence from the preceding W-day window:

$$\mathbf{s}_i = \{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \ldots, \mathbf{e}_{i,T_i}\}, \tag{1}$$

where each event $\mathbf{e}_{i,j}$ corresponds to a booking or refund action, and T_i denotes the number of observed events within the window, which varies across users.

The objective is to learn a detection model F_θ that predicts whether the tickets booked within the W-day window will collectively result in bulk refunds:

$$y_i = F_\theta(\mathbf{s}_i), \tag{2}$$

where ($y_i \in \{0, 1\}$) indicates BRB risk ($y_i = 1$) or normal behavior ($y_i = 0$). Beyond risk prediction, BulkShield identifies key behavioral evidence from the sequence and generates natural-language explanations using LLM that summarizes booking context and inferred intent for operational decision-making. An overview of this process is shown in Fig. 2.

4 Methodology

We introduce *BulkShield*, an end-to-end framework for proactive and interpretable Bulk Refund Booking detection. The framework consists of four components:

- **Risk-Aware Unsupervised BRB Pseudo-Labeling (Sec. 4.1)** We derive BRB pseudo-labels without predefined risk criteria by clustering the subsequent refund outcomes of tickets booked within a W-day window, and use them as reference supervision for booking-time risk prediction.
- **Expert-Guided Sequence Feature Extraction (Sec. 4.2).** Guided by domain expert input, we extract structured sequential features from user actions observable before prediction, capturing practical signals of booking context bulk-refund misuse.
- **Transformer-Based BRB Detection (Sec. 4.3).** We model booking–refund event sequences with a Transformer-based classifier to capture long-range and recurring patterns for booking-time BRB risk prediction.
- **Attention-Guided LLM Explanation (Sec. 4.4).** We select high-attention events as behavioral evidence and prompt an LLM to generate practitioner-oriented natural-language explanations that support operational decision-making.

4.1 Risk-Aware Unsupervised Pseudo-Labeling

We propose an unsupervised pseudo-labeling framework that selects an observation window where refund risk is reliably distinguishable (Sec. 4.1.1) and then derives BRB risk groups from post-hoc refund outcomes within that window (Sec. 4.1.2).

4.1.1 Optimal Observation Window Selection. For proactive BRB detection, it is important to determine how far back a user's booking history should be observed. Bookings within this window are later summarized by their refund outcomes, which provide the basis for identifying whether meaningful BRB risk patterns have emerged [21]. Too-short windows may miss accumulated refund patterns, whereas overly long windows may mix outdated history with current behavior, making pseudo-labels less representative of the user's booking-time risk state [7].

We therefore select the shortest window that yields both (i) clear separation between potential BRB-risk and normal behavior and (ii) stable discrimination across repeated sampling [21]. Given a user set $\mathcal{U} = \{1, \ldots, N\}$, we randomly sample an inference booking date T_i for each user i. For each candidate window length W, we collect tickets booked during the preceding W days and summarize their subsequent refund outcomes as

$$\mathbf{r}_i(W) = (\text{count}_i(W),\ \text{rate}_i(W),\ \text{amount}_i(W)), \tag{3}$$

where *count* is the number of subsequently refunded tickets, *rate* is their ratio among tickets booked within (W), and *amount* is the total corresponding refund amount.

For each window length W, the refund summaries $\{\mathbf{r}_i(W)\}_{i=1}^{N}$ are highly imbalanced: most users exhibit low refund activity, while only a small subset shows elevated refund behavior. However, intermediate cases appear between clearly normal users and users with consistently high refund counts, rates, and amounts, creating uncertainty near the normal–risk boundary. This continuous and overlapping structure makes fixed thresholds or hard clustering methods (*e.g.*, k-means [35], etc) unsuitable, as they force users into rigid groups despite uncertainty near the boundary. We therefore adopt a Gaussian Mixture Model (GMM) [1, 24],

$$p(\mathbf{r}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{r} \mid \boldsymbol{\mu}_k, \Sigma_k), \tag{4}$$

which captures latent refund-behavior groups through soft probabilistic assignments, with the number of groups K selected via BIC [32]. Each sample is assigned to the component with the largest posterior probability; the component with the largest median refund amount is treated as the potential BRB risk group, and the remaining components are aggregated as the normal group for binary supervision.

For each (W), we quantify the separability between the risk and normal groups using the Bhattacharyya distance [3],

$$D_B(W) = \frac{1}{8}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_n)^\top \Sigma^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_n) + \frac{1}{2} \ln \frac{|\Sigma|}{\sqrt{|\Sigma_r||\Sigma_n|}}, \tag{5}$$

which captures both mean separation and distributional overlap between the two groups. To assess robustness, we repeat this process M times with different randomly sampled booking times and obtain a distribution of $D_B(W)$ for each window length.

We select the optimal observation window as the shortest W that achieves both sufficiently large average separation and low variability across repetition [21]:

$$W_{\text{opt}} = \arg\min_{W} W \quad \text{s.t.} \quad \mathbb{E}[D_B(W)] \geq \tau,\ \text{Var}[D_B(W)] \leq \epsilon. \tag{6}$$

This criterion ensures that the shortest window in which refund behavior groups are both separable and stable, enabling reliable BRB pseudo-label generation.

4.1.2 Bulk Refund Booking Pseudo-Label Generator. With W_{opt} fixed, we fit a GMM to the refund summary vectors $\mathbf{r}_i(W_{\text{opt}})$ and define the component with the largest median refund amount as the latent BRB risk group C_{risk}. Each booking-triggered user sample is then assigned a binary pseudo-label:

$$y_i^{\text{pseudo}} = \begin{cases} 1, & \text{if } \mathbf{r}_i(W_{\text{opt}}) \in C_{\text{risk}}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

These pseudo-labels serve as reference supervision for training and evaluating proactive BRB detection models at booking time.

4.2 Expert-Guided Sequence Feature Extraction

To avoid future information leakage, we extract features only from time-ordered booking and refund events observed before each prediction point. These features are aligned with pseudo-labels derived from the subsequent refund outcomes of tickets booked within the W_{opt}-day window. Guided by domain experts, we focus on sequential patterns that reflect booking context and BRB-related suspicious signals rather than isolated transactions.

We organize the extracted features into three complementary categories: (1) **Ticketing Attributes**, covering transaction-level context such as seat counts, monetary amounts, lead times, and holding durations; (2) **Usage Intention**, capturing route repetition, travel directionality, and temporal concentration that characterize recurring travel behavior; and (3) **Suspicious Signals**, representing expert-informed risk indicators such as overlapping active tickets, repeated same-route bookings, and concentrated refund activity. A complete list of the 32 extracted features, together with their distributions (mean ± std), is provided in Appendix 4, Table B.

4.3 Transformer-Based BRB Detection

We employ a Transformer [42]-based sequence classifier for proactive BRB risk detection. BRB risk emerges from long-range, repetitive, and temporally concentrated booking–refund interactions that are individually permissible but become risky when accumulated. Thus, unlike anomaly detection methods [2, 31, 45] that model risk as deviations from normal behavior, historical-statistics-based classifiers that discard event order, or recurrent models that rely on recursive state propagation, the Transformer is well suited for directly modeling interactions among temporally distant events through self-attention.

4.3.1 Input Representation. For each user i, we construct a time-ordered event sequence $\mathbf{s}_i = \{\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,T_i}\}$, where each event corresponds to a booking or refund action observed before the prediction time within a rolling window of W days. Each event is represented by categorical or ordinal features $\mathbf{c}_{i,j}$, numerical or binary features $\mathbf{n}_{i,j}$, and the inter-event time gap $\Delta_{i,j}$. Categorical

and ordinal features are embedded, numerical and binary features are log-transformed and linearly projected, and time gaps are discretized and embedded to capture temporal concentration patterns. The final event embedding is defined as

$$\mathbf{z}_{i,j} = \mathrm{Emb}_{\mathrm{cat}}(c_{i,j}) + \mathrm{Proj}_{\mathrm{num}}(n_{i,j}) + \mathrm{Emb}_{\Delta}(\Delta_{i,j}),$$

with padding and attention masks applied for variable-length sequences.

4.3.2 Sequence Modeling. We prepend a learnable [CLS] token to each event sequence and add positional encodings to preserve event order [8]. An L-layer Transformer encoder with masked self-attention then produces the final [CLS] representation $\mathbf{h}_i^{\mathrm{cls}}$, which summarizes the user's accumulated booking–refund behavior by integrating repeated and temporally distant events relevant to BRB risk.

4.3.3 Output Layer. The [CLS] representation, $\mathbf{h}_i^{\mathrm{cls}}$, is passed to a classification head to estimate BRB risk:

$$\hat{\mathbf{y}}_i = \mathrm{Head}\left(\mathbf{h}_i^{\mathrm{cls}}\right) \in \mathbb{R}^2,$$

which is transformed via softmax into a BRB risk probability. This enables BulkShield to proactively identify high-risk users from sequential booking-time behavior, rather than relying on static post-hoc refund thresholds.

4.3.4 The Optimization Process. We train the model using binary cross-entropy loss [23], with pseudo-labels derived from post-hoc refund outcomes serving as reference supervision. Since the input sequence contains only events observed before prediction time, the model learns early booking-time behavioral signals associated with subsequent bulk-refund risk.

4.4 Attention-Guided LLM Explanation

We generate evidence-based operational explanations for users flagged as BRB-risky. Given a trained Transformer classifier, BulkShield uses model attention to identify explanation-critical booking–refund events and prompts an LLM to summarize BRB risk evidence and inferred booking intent.

4.4.1 Attention-Guided Prompt Engineering. For each user sequence, we extract attention weights from the final Transformer encoder layer and compute head-averaged attention from the [CLS] token to each event token [33]. This identifies prediction-critical booking–refund events that ground the subsequent explanation. Let $\mathbf{A} \in \mathbb{R}^{H\times S\times S}$ denote the multi-head attention matrix, where $S = T + 1$. The importance score of event j is computed as

$$s_j = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}_{h,\mathrm{CLS},j}, \quad j \in \{1,\ldots,T\}. \tag{8}$$

Based on these scores, we first select the Top-K event indices and then retrieve the corresponding booking–refund events as salient evidence for explanation:

$$\mathcal{I}_{\text{Top-}K} = \text{Top-}k_{j\in\{1,\ldots,T\}}(s_j), \qquad \mathcal{E}_{\text{Top-}K} = \{\mathbf{e}_{i,j} \mid j \in \mathcal{I}_{\text{Top-}K}\}. \tag{9}$$

Using this evidence, we construct a constrained prompt [43] that guides the LLM to generate explanations grounded in selected booking–refund behaviors:

$$P_{E\to X} = P_{\mathrm{Role}} + P_{\mathrm{Evidence}} + P_{\mathrm{Reason}} + P_{\mathrm{Explain}}. \tag{10}$$

Here, P_{Role} defines the LLM as a railway-domain analyst, P_{Evidence} provides the selected evidence events, P_{Reason} guides evidence-grounded reasoning, and P_{Explain} enforces a fixed output structure. Illustrative prompt examples are provided in Appendix C.

4.4.2 LLM-Assisted Explanation Generation. Given the constructed prompt, we generate operational explanations using an open-source LLM, such as Llama-3.1-70B-Instruct [10], to support privacy-preserving deployment within controlled railway infrastructure. The generated explanation summarizes both BRB risk evidence and plausible booking intent, enabling practitioners to review high-risk users with concrete event-level support rather than risk scores alone.

5 Experiments

This section evaluates BulkShield by addressing the following research questions:

- **RQ1.** Do unsupervised risk-aware pseudo-labels align with domain experts' practical boundaries between normal and risky behavior? (Sec. 5.3)
- **RQ2.** How effectively does BulkShield proactively detect BRB risk compared to baseline methods? (Sec. 5.4)
- **RQ3.** Do the generated explanations provide actionable and trustworthy support for real-world decision-making? (Sec. 5.5)

5.1 Dataset

5.1.1 Dataset Description. We use a large-scale SRT ticketing dataset containing ticket-level booking and refund records from 6,756,549 customers. The dataset covers train journeys departing between January 1 and December 31, 2024, booked through mobile and online platforms. Each record includes 148 attributes covering (1) customer and booking information (e.g., anonymized customer IDs, booking types, ticket status, and purchase timestamps), (2) train and journey details (e.g., origin–destination station codes, travel times, and seat attributes), and (3) payment, cancellation, and refund information (e.g., fares, discounts, refund fees, and refund records).

5.1.2 Data Preprocessing. To ensure data reliability, we applied two filtering steps. First, we retained users with at least one booking and an observed usage period of at least 60 days, excluding one-time or short-term users. Second, we removed records with invalid departure or arrival timestamps (e.g., 00:00:00 or 23:59:59) or implausible departure–arrival gaps exceeding two days. The final dataset contains 1,697,500 users, with an average of 11.45 active days and 33.56 events per user, and is randomly split into training (1,357,991) and test (339,509) sets.

5.1.3 Sample Construction. We construct one individual-level sample for each passenger at an inference booking time t. Input features are extracted only from booking–refund events observed within the past window $[t - W_{\mathrm{opt}}, t]$. Pseudo-labels are assigned from the final refund outcomes of tickets booked within this window, including refunds already observable before t and subsequent refunds within $[t, t + \alpha]$. To preserve the proactive detection setting, we retain

only samples whose BRB pseudo-label is determined by refund outcomes occurring after t. Thus, future refund outcomes define the window-level label but are never included as booking-time input features.

The procedure consists of the following steps.

(1) **Observation window selection:** Using only training passengers, we determine W_{opt} via Sec. 4.1.1 and fix it for all experiments.

(2) **Inference date selection:** For each passenger, we select one booking time t as the prediction point, with the preceding W_{opt} days used for feature extraction.

(3) **Pseudo-label generation:** The unsupervised pseudo-label generator is fitted on training samples to identify the risk group C_{risk}, which is then fixed and applied to test samples.

(4) **Feature extraction:** Sequence features are extracted from events within the observation window for both training and test samples. In the training set, all numerical and binary features show significant differences between risk and normal groups ($p < 0.05$) in either mean or standard deviation [26].

5.2 Experimental Setups

5.2.1 Training Details. We use a Transformer classifier with 3 encoder layers and 4 attention heads, where $d_{\text{model}} = 128$ and $d_{\text{ff}} = 256$, and apply a dropout rate of 0.1. The model is trained in PyTorch on an NVIDIA H100 GPU with a batch size of 128, learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, and a maximum of 20 epochs. For consistency, baseline models use default library hyperparameters unless otherwise specified. Each experiment is repeated 20 times with different random seeds, and we report the mean and standard deviation.

5.2.2 Evaluation Metric. We evaluate BulkShield using metrics aligned with each research question.

- **RQ1.** Pseudo-label validity is measured by Cohen's κ [18], assessing agreement between learned risk labels and expert judgments beyond chance.
- **RQ2.** Proactive BRB detection performance is evaluated using Macro-average Precision, Recall, and F1-score to account for class imbalance.
- **RQ3.** Explanation usefulness is evaluated through expert-based quantitative and qualitative assessments, focusing on actionability and trustworthiness.

5.3 Pseudo-Label Validation by Experts (RQ1)

This section validates the BRB pseudo-labels by examining whether the learned risk groups are separable and stable across observation windows (Sec. 5.3.1), and whether they align with practitioner judgment in a structured human-in-the-loop study (Sec. 5.3.2).

5.3.1 Optimal Window Size Selection. Fig. 3 shows how the observation window length affects the separability and stability of refund behavior groups. We evaluate window sizes up to $W_{\max} = 60$ days to cover both short- and long-term behavioral patterns.

- **Early regime ($W < 20$ days):** Short windows provide insufficient post-hoc refund information, resulting in low and unstable separability between risk and normal groups.
- **Transition to optimal regime ($W \approx 28$ days):** As W increases, separability improves and variability decreases. At $W = 28$ days (4 weeks), the mean Bhattacharyya distance stabilizes and variance reaches its minimum, indicating robust group separation.
- **Saturation regime ($W > 28$ days):** Longer windows provide little additional separability and slightly increase variability, suggesting the inclusion of redundant or noisy historical information.

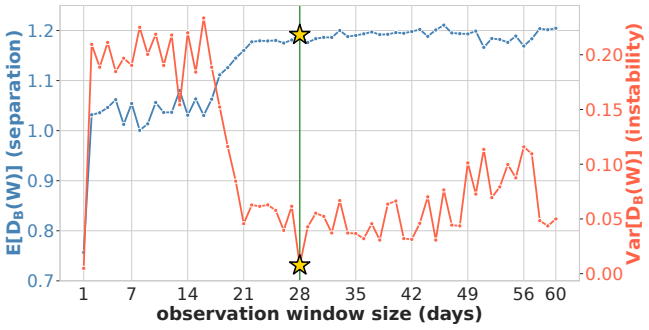


Figure 3: Effect of observation window length on the separability and stability of BRB pseudo-label groups. The expected Bhattacharyya distance, $E[D_B(W)]$ (blue), captures the average separation between risk and normal group, while the variance, $Var[D_B(W)]$ (red), measures instability across repeated sampling. The optimal window ($W_{\text{opt}} = 28$) is selected as the shortest window achieving stable and well-separated pseudo-label groups.

Based on these results, we set $W_{\text{opt}} = 28$ days as the final observation window, as it is the shortest window achieving stable and well-separated refund behavior groups. This also aligns with SR's one-month advance reservation policy, reflecting the practical temporal scope in which BRB behaviors emerge.

5.3.2 Pseudo-Labels validation from expert. After fixing W_{opt}, we trained the unsupervised pseudo-label generator on the training set and applied the learned risk-group definition to both training and test sets. As shown in Table 1, the resulting pseudo-labels show statistically significant separation between risk and normal groups across refund-related patterns, with $p < 0.05$ [26].

To assess practical validity, we conducted a human-in-the-loop study with eight internally recruited SR practitioners, whose profiles are provided in Appendix A. Considering class imbalance, we constructed a stratified set of 30 test cases: 5 extreme rule-violation case that exceeded existing rule-based refund thresholds as attention checks, 15 pseudo-labeled BRB cases, and 10 normal cases. Experts reviewed standardized post-hoc refund summaries for tickets booked within W_{opt}, including booking count, refund count, refund amount, and refund ratio, without access to model outputs or clustering results. They independently labeled each case as Definite BRB (◯), Suspected BRB (△), or Normal (×), with Definite and Suspected BRB treated as positive for agreement analysis. Since all experts passed the attention checks, all responses were retained.

The pseudo-labels show strong agreement with expert judgments, with $\kappa = 0.8862$ [18], indicating that the unsupervised

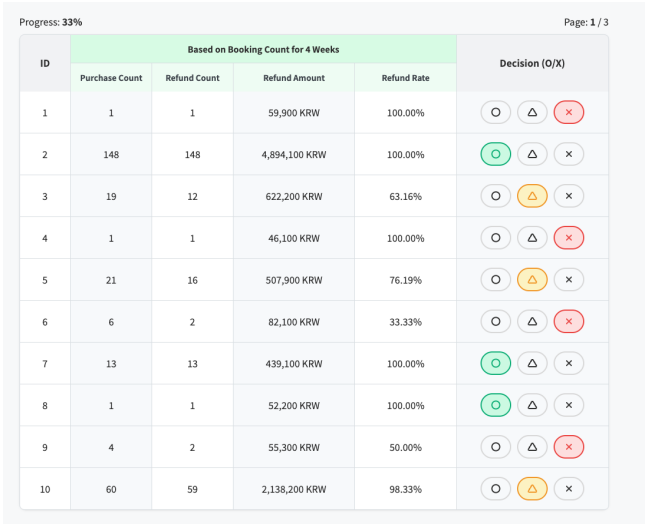


Figure 4: Survey interface for pseudo-label validation. Experts classified each case based on standardized post-hoc refund statistics computed for tickets booked within W_{opt}, while model outputs and clustering results were hidden to ensure blinded judgment.

risk groups align with expert-recognized problematic refund patterns and can serve as reference supervision for BulkShield. In contrast, the current fixed rule-based criteria (refund amount > KRW 1,000,000 and refund rate > 90%) produce sparse labels, 0.036%, and much lower expert agreement, $\kappa = 0.2824$, suggesting that coarse thresholds fail to capture heterogeneous and overlapping BRB risk patterns for proactive detection.

5.4 Overall BRB Detection Performance (RQ2)

We evaluate the proactive BRB detection performance of BulkShield against three groups of baselines: (i) anomaly detection methods that model BRB as deviations from normal behavior (Deep SVDD, LSTM Autoencoder, USAD), (ii) historical statistics-based classifiers using aggregated refund features within the observation window (Logistic Regression, SVR, MLP), and (iii) sequential models that capture temporal dependencies in booking–refund sequences (RNN, LSTM, GRU).

5.4.1 Performance Evaluation. Table 2 summarizes the detection results. Anomaly detection methods show limited performance, with Macro F1 scores of 0.6021–0.6809, suggesting that BRB is not well captured as a simple deviation from normal behavior. Historical statistics-based classifiers improve performance by using aggregated refund features, while sequential baselines perform better by modeling temporal dependencies.

BulkShield achieves the best overall performance, with a Macro F1 of 0.8909, outperforming the strongest sequential baseline, GRU, with 0.8627. This indicates that BulkShield effectively captures long-range and repetitive booking–refund patterns while maintaining balanced precision and recall.

Table 1: Statistical comparison [26] of post-hoc refund outcomes for bookings within W_{opt} across risk and normal groups. Train: n_{train} = 1,357,991 (n_{risk} = 32,774; 2.41%), Test: n_{test} = 339,509 (n_{risk} = 8,078; 2.38%).

Metric	Training Set			Test Set		
	Risk	Normal	p-value	Risk	Normal	p-value
Refund rate	0.783	0.333	< 0.05	0.786	0.333	< 0.05
Refund amount (KRW)	492,100	22,700	< 0.05	499,650	22,700	< 0.05
Refund count	13	1	< 0.05	13	1	< 0.05

Table 2: BRB detection performance compared with baseline models. The best results are highlighted in bold.

Model	Macro F1	Precision	Recall
Deep SVDD	0.6309 ± 0.018	0.6405 ± 0.021	0.6227 ± 0.019
LSTM Autoencoder	0.6809 ± 0.015	0.6703 ± 0.017	0.6930 ± 0.020
USAD	0.6021 ± 0.014	0.5775 ± 0.018	0.6648 ± 0.022
Logistic Regression	0.7890 ± 0.002	0.8507 ± 0.018	0.7753 ± 0.013
SVR	0.7537 ± 0.042	0.7996 ± 0.055	0.7369 ± 0.087
MLP	0.5032 ± 0.090	0.8596 ± 0.028	0.5046 ± 0.060
RNN	0.8597 ± 0.001	0.8736 ± 0.010	0.8472 ± 0.010
LSTM	0.8180 ± 0.009	0.8442 ± 0.013	0.8020 ± 0.019
GRU	0.8627 ± 0.002	0.8719 ± 0.009	0.8520 ± 0.008
BulkShield (Ours)	**0.8909 ± 0.002**	**0.8852 ± 0.002**	**0.8985 ± 0.012**

5.4.2 Ablation Study on Expert-Guided Features. To assess the contribution of expert-guided feature design, we remove each feature group introduced in Sec. 4.2. Removing Ticketing Attributes, Usage Intention, and Suspicious Signals reduces Macro F1 from 0.8909 to 0.8692, 0.8818, and 0.8739, respectively. These results indicate that the three feature groups provide complementary signals for BRB detection.

5.4.3 Temporal Validation for Deployment Realism. To assess temporal robustness, we train BulkShield on January–October 2024 data and test it on November–December 2024 data. The model achieves a Macro F1 of 0.9036, comparable to the random split result of 0.8909. This suggests that BulkShield captures stable BRB behavior patterns beyond a specific random split, supporting its applicability in deployment-oriented settings.

5.5 Expert Validation of Explanations (RQ3)

To evaluate the practical value of LLM-assisted explanations, we conducted an expert validation study with the SR practitioners who participated in the pseudo-label validation study (Sec. 5.3.2). Experts reviewed true-positive test cases, where samples were pseudo-labeled as BRB and correctly predicted as BRB by BulkShield before the label-defining refund outcome occurred. Each case was presented with an LLM-generated explanation and visual summaries of booking–refund behavior (Figs. 5 and 6).

The evaluation was designed to assess operational usefulness rather than predictive accuracy. For the quantitative assessment, experts rated the explanations in terms of interpretability (Q1), relevance (Q2), trustworthiness (Q3), operational utility (Q4), misinterpretation risk (Q5), and adoption intention (Q6), as shown in

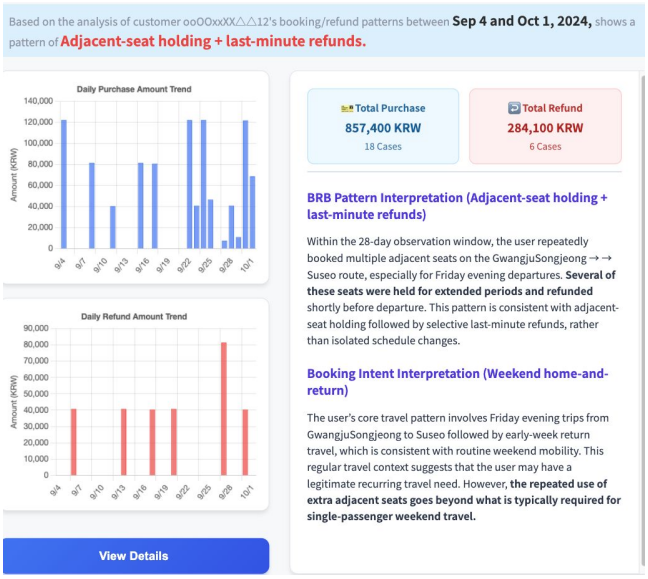


Figure 5: Initial diagnostic view of the expert dashboard for LLM-assisted BRB explanation validation. The top panel summarizes the analysis period and predicted BRB pattern, the left panel visualizes aggregated daily booking–refund trends within W_{opt}, and the right panel provides structured LLM explanations of BRB risk and booking intent. Experts can select *View Details* in the left panel to inspect the attention-guided evidence events shown in Fig. 6.

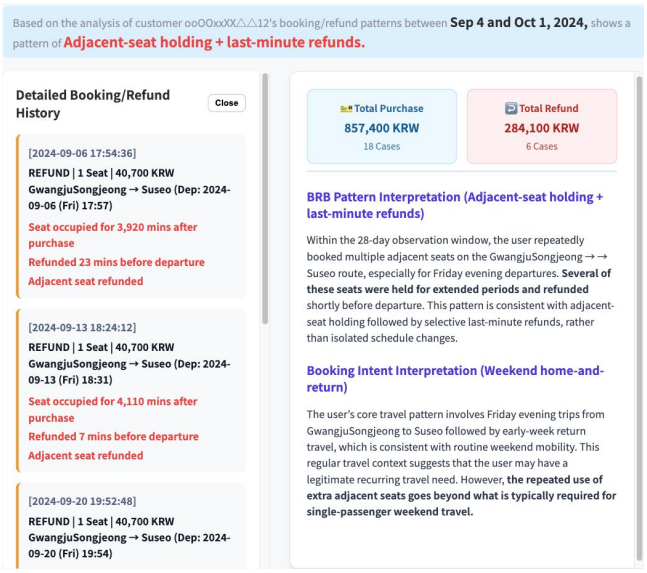


Figure 6: Detailed inspection view accessed through *View Details* in Fig. 5. The view presents attention-guided Top-K evidence events and corresponding transaction logs, highlighting key risk indicators such as prolonged seat occupation, repeated same-route refunds, and last-minute cancellations to support evidence-grounded validation.

Fig. 7. We also collected qualitative feedback through open-ended questions on how booking-intent explanations influenced expert judgment and how visual risk presentation could be improved.

5.5.1 Quantitative Results. The explanations received consistently high ratings for interpretability, relevance, trustworthiness, operational utility, and adoption intention (Q1–Q4, Q6), with all mean scores above 4.5 on a five-point Likert scale. The reverse-coded item on misinterpretation risk (Q5) showed a low mean score, indicating that experts generally did not find the explanations confusing or misleading. These results suggest that the explanations were perceived as useful for practitioner review.

5.5.2 Qualitative Results. Qualitative feedback further revealed key strengths and limitations of the explanation framework.

Strengths. Experts (P3, P8) noted that the explanations helped distinguish heterogeneous refund intentions rather than treating bulk refunds as a single uniform behavior, supporting more nuanced operational judgment. Others (P1, P6) emphasized that time-series and graphical summaries improved interpretability compared with inspecting raw transaction logs.

Limitations and Improvement Suggestions. Experts highlighted the need for richer contextual cues. In particular, aggregating events only by transaction date can obscure important information, such as the actual travel dates of refunded tickets (P4, P6). P3 also suggested finer-grained intent categories to better capture diverse and less common refund behaviors. These findings indicate that future explanation systems should incorporate richer travel-date context and more detailed booking-context categories.

6 Discussion

6.1 BulkShield for Smart Railway Operations

When deployed in real-world SR booking systems, BulkShield can help address structural inefficiencies by supporting proactive and interpretable BRB risk management. We summarize its implications from three stakeholder perspectives:

- **User perspective: improved seat accessibility and fairness.** By identifying strategic bulk seat holding earlier, BulkShield alleviates artificial seat shortages, improving ticket availability and perceived fairness, particularly during peak travel periods.
- **Operator perspective: timely intervention and seat recovery.** Early BRB risk identification can support practitioner review before refunds accumulate at scale, creating opportunities for timely intervention, seat recovery, and revenue-loss risk mitigation.
- **System-level perspective: behavior-aware and sustainable mobility management.** By linking BRB risk to underlying booking and mobility behaviors, BulkShield supports proactive and behavior-aware management of shared seat capacity, aligning operational decisions with efficiency, equity, and public trust.

Overall, these implications show how behavior-aware AI can shift smart railway operations from reactive enforcement to proactive, interpretable, and policy-aligned mobility management.

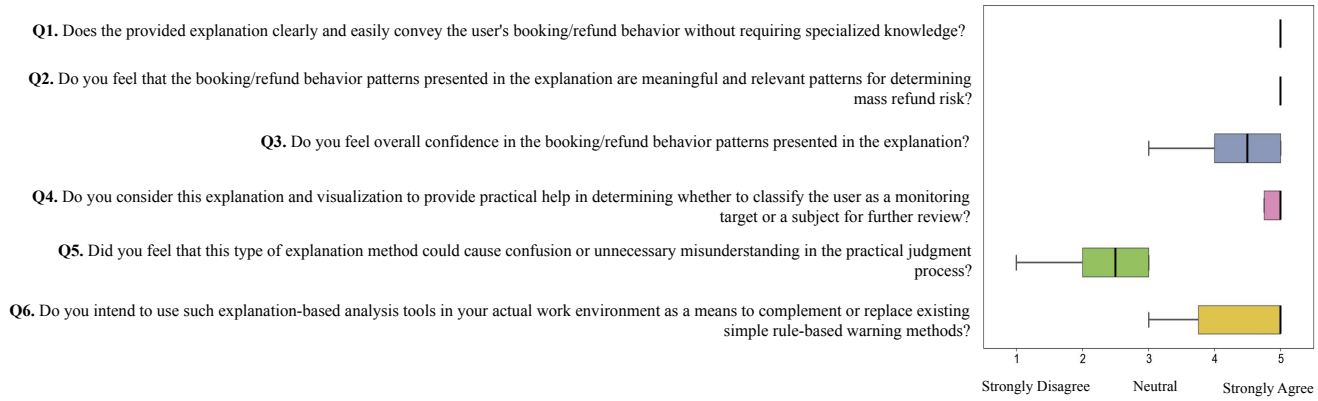


Figure 7: Expert ratings of LLM-assisted BRB explanations across six evaluation questions (Q1–Q6) on a five-point Likert scale. Boxes indicate the interquartile range (IQR), and central lines indicate medians.

6.2 Limitations and Future Work

Despite its practical implications, this work has several limitations.

First, BulkShield relies on predefined BRB pattern types and booking-context categories derived from domain-expert knowledge. While these definitions support interpretable and operationally meaningful explanations, booking and refund strategies may evolve as users adapt to policy or monitoring practices, and emerging behaviors may not be fully captured by the current categories.

Second, although expert feedback supports the usefulness of LLM-generated explanations, the framework does not yet include a formal mechanism for verifying their factual correctness, completeness, or objectivity. Without structured validation and human oversight, explanations may remain incomplete or misleading in operational settings.

To address these challenges, future work will focus on improving both adaptability and governance. We plan to refine behavioral definitions through continuous collaboration with domain experts and explore adaptive, data-driven methods to identify previously unseen BRB patterns. In parallel, we aim to incorporate human-in-the-loop validation processes in which LLM-generated explanations are systematically reviewed and contextualized before informing enforcement or policy decisions.

6.3 Ethical Considerations

This study was conducted under strict ethical and legal safeguards to protect passenger privacy and prevent misuse.

Legal basis and purpose limitation. All analyses were conducted under a formal pseudonymized data processing agreement between SR and the research institution, which defined the purpose, scope, and duration of data use. The data were used solely for research on ticketing behavior and BRB prevention, in compliance with the Personal Information Protection Act of Korea and related regulations.

Privacy protection and data security. Passenger identifiers were pseudonymized and de-identified before analysis, and no raw personal identifiers were accessible to the research team. All datasets were encrypted using AES-256-CBC and stored on secure servers with controlled access.

Governance and accountability. All researchers signed confidentiality and data security agreements prohibiting unauthorized disclosure, secondary use, re-identification, or third-party sharing. Data processing activities were logged and subject to oversight by the data provider, and all pseudonymized data are mandatorily destroyed or returned upon contract completion.

IRB consideration. This study did not require IRB approval, as it involved no direct interaction with human subjects and relied exclusively on pseudonymized administrative records processed under a formal data use agreement.

7 Conclusion

In this study, we characterize Bulk Refund Booking as a system-level demand distortion phenomenon driven by cumulative and sequential booking–refund behaviors, rather than static post-hoc misuse. To study and address this phenomenon, we introduce *BulkShield*, an end-to-end framework that combines unsupervised risk labeling, sequence-based BRB detection, and LLM-assisted explanations to capture how behavioral risk accumulates over time. Expert evaluations confirm that the inferred risk boundaries align with domain judgment, supporting their validity as reference supervision for proactive detection. Furthermore, the generated explanations enable intent-aware interpretation of booking behavior, providing transparent evidence to inform policy interventions in public transportation systems. Overall, this work shows that behavior-aware and interpretable AI can uncover demand distortion mechanisms in railway ticketing systems, informing proactive and equitable capacity management.

Acknowledgments

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Climate, Energy & Environment (MCEE) of the Republic of Korea (Nos. RS-2025-02654073 and RS-2026-25530097).

References

[1] Anas Al-Lahham, Nurbek Tastan, Muhammad Zaigham Zaheer, and Karthik Nandakumar. 2024. A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6793–6802.

[2] Julien Audibert, Pietro Michiardi, Fr'ed'eric Guyard, S'ebastien Marti, and Maria A. Zuluaga. 2020. USAD: Unsupervised Anomaly Detection on Multivariate Time Series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. 3395–3404.

[3] Anil Kumar Bhattacharyya. 1943. On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* 35 (1943), 99–109.

[4] Richard J. Bolton and David J. Hand. 2002. Statistical Fraud Detection: A Review. *Statist. Sci.* 17, 3 (2002), 235–255.

[5] Bukyung Maeil. 2025. SRT No-show Seats and Refund Trends. http://www.bukn.kr/news/articleView.html?idxno=26268 Accessed: 2026-06-09.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *Comput. Surveys* 41, 3 (2009).

[7] Weiwei Deng, Xiaoliang Ling, Yang Qi, Tunzi Tan, Eren Manavoglu, and Qi Zhang. 2018. Ad Click Prediction in Sequence with Long Short-Term Memory Networks: an Externality-aware Model. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1065–1068. doi:10.1145/3209978.3210071

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[9] Dong-A Ilbo. 2024. Minimal Penalties Collected from Large-scale Refund Users. https://www.donga.com/news/Economy/article/all/20241003/130149251/1 Accessed: 2026-06-09.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018), 1–42.

[12] Hankyoreh. 2024. Bulk Ticket Bookings and Refunds Lead to Railway Seat Hoarding and Resale. https://www.hani.co.kr/arti/society/society_general/1220551.html Accessed: 2026-02-05.

[13] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5549–5581.

[14] Mengda Huang, Yang Liu, Xiang Ao, Kuan Li, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. 2022. AUC-oriented Graph Neural Network for Fraud Detection. In *Proceedings of the ACM Web Conference 2022*. 1311–1321.

[15] Sujeong Hur, Yura Lee, Joongheum Park, Yeong Jeong Jeon, Jong Ho Cho, Duck Cho, Dobin Lim, Wonil Hwang, Won Chul Cha, and Junsang Yoo. 2025. Comparison of SHAP and Clinician Friendly Explanations Reveals Effects on Clinical Decision Behaviour. *NPJ Digital Medicine* 8, 1 (2025), 578.

[16] KBS News. 2024. Public Complaints Regarding SRT Seat Availability. https://news.kbs.co.kr/news/mobile/view/view.do?ncd=8427574 Accessed: 2026-06-09.

[17] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based Explanations as Feature Importance Measures. In *International Conference on Machine Learning*. PMLR, 5491–5500.

[18] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* (1977), 159–174.

[19] Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. 2021. Pick and Choose: A GNN-based Imbalanced Learning Approach for Fraud Detection. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3168–3177. doi:10.1145/3442381.3449989

[20] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[21] Tianyang Luo, Xikun Zhang, and Dongjin Song. 2026. FTSCommDetector: Discovering Behavioral Communities through Temporal Synchronization. arXiv:2510.00014 [cs.SI] https://arxiv.org/abs/2510.00014

[22] Xiaoxiao Ma, Ruikun Li, Fanzhen Liu, Kaize Ding, Jian Yang, and Jia Wu. 2024. Graph Anomaly Detection with Few Labels: A Data-Centric Approach. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2153–2164.

[23] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 23803–23828.

[24] Geoffrey J McLachlan and Suren Rathnayake. 2014. On the Number of Components in a Gaussian Mixture Model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4, 5 (2014), 341–355.

[25] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38.

[26] Nadim Nachar et al. 2008. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology* 4, 1 (2008), 13–20.

[27] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. 2010. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *arXiv preprint arXiv:1009.6119* (2010).

[28] Railway Daily. 2023. Exploiting Full Ticket Refunds to Accumulate Credit Card Reward Points. https://www.redaily.co.kr/news/articleView.html?idxno=4374 Accessed: 2026-02-05.

[29] Railway Daily. 2024. Seat Shortages Despite Existing Capacity. https://www.redaily.co.kr/news/articleView.html?idxno=12497 Accessed: 2026-06-09.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[31] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 4393–4402. https://proceedings.mlr.press/v80/ruff18a.html

[32] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.

[33] Sofia Serrano and Noah A Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2931–2951.

[34] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034* (2013).

[35] Kristina P. Sinaga and Miin-Shen Yang. 2020. Unsupervised K-Means Clustering Algorithm. *IEEE Access* 8 (2020), 80716–80727. doi:10.1109/ACCESS.2020.2988796

[36] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel. *Nature Machine Intelligence* 5, 8 (2023), 873–883.

[37] SR Corporation. 2025. Introduction to SRT Services. https://www.srail.or.kr Accessed: 2026-06-09.

[38] SR Corporation. 2025. Refund Policy and Sanctions for Excessive Refunds. https://etk.srail.kr/cms/article/view.do?postNo=737&pageId=TK0502000000 Accessed: 2026-06-09.

[39] SR Corporation. 2025. SRT Mobile Application and Online Ticketing Services. https://etk.srail.co.kr Accessed: 2026-06-09.

[40] SR Corporation. 2025. SRT Passenger Growth Statistics. https://www.srail.or.kr/cms/article/view.do?postNo=1529&pageId=KR0502000000 Accessed: 2026-06-09.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).

[43] Xuezhi Wang and Denny Zhou. 2024. Chain-of-Thought Reasoning without Prompting. *Advances in Neural Information Processing Systems* 37 (2024), 66383–66409.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[45] Yuanyuan Wei, Julian Jang-Jaccard, Wen Xu, Fariza Sabrina, Seyit Camtepe, and Mikael Boulic. 2023. LSTM-Autoencoder-based Anomaly Detection for Indoor Air Quality Time-Series Data. *IEEE Sensors Journal* 23, 4 (2023), 3787–3800.

[46] Chengdong Yang, Hongrui Liu, Daixin Wang, Zhiqiang Zhang, Cheng Yang, and Chuan Shi. 2025. FLAG: Fraud Detection with LLM-enhanced Graph Neural Network. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 5150–5160. doi:10.1145/3711896.3737220

[47] Ge Zhang, Zhao Li, Jiaming Huang, Jia Wu, Chuan Zhou, Jian Yang, and Jianliang Gao. 2022. eFraudCom: An E-commerce Fraud Detection System via Competitive Graph Neural Networks. *ACM Trans. Inf. Syst.* 40, 3, Article 47 (March 2022), 29 pages. doi:10.1145/3474379

A Participant Demographics and Recruitment

As summarized in Table 3, we recruited eight SR practitioners and divided them into two complementary groups to reflect both institutional policy judgment and operational data validation.

- **Group 1: Strategy & Policy (P1–P4).** This group includes managerial-level personnel, such as the General Manager of Sales Planning and the Head of Marketing. They assessed whether the observed behaviors could be interpreted as policy or regulatory violations from an institutional enforcement perspective.
- **Group 2: Data & Operations (P5–P8).** This group includes data analysts, membership managers, and administrative practitioners. They evaluated whether the observed data patterns were operationally anomalous and consistent with known misuse behaviors.

Participant	Age	Degree	Gender	Job Role
P1	47	Bachelor's	Male	Sales Planning (Overall Operations)
P2	47	Bachelor's	Male	Marketing Operations
P3	39	Bachelor's	Male	Passenger Policy Management
P4	35	Bachelor's	Male	Headquarters Staff
P5	31	Bachelor's	Male	Administrative Operations
P6	33	Bachelor's	Female	Data Analysis
P7	28	Bachelor's	Female	Membership Policy
P8	31	Bachelor's	Male	Administrative Support

Table 3: Expert profiles for the human-in-the-loop evaluation

B Expert-Guided Feature Extraction

This section describes 32 expert-guided sequential features that characterize ticketing behavior across booking and refund events. As summarized in Table 4, the feature set comprises 16 *Ticketing Attributes*, 8 *Usage Intention* features, and 8 *Suspicious Signals*. For numerical and binary features, we additionally report their aggregated distributions over the observation window to provide an overall summary of feature behavior.

C Attention-Guided Prompt Structure

This section presents a structured prompting strategy for explaining BRB risk from attention-selected evidence.

C.1 P_{Role}

```
 You are an analyst supporting a railway operator's
enforcement decisions for suspicious ticketing behavior.
Write an operationally useful, evidence-based explanation.

Rules:
- Think step-by-step internally; DO NOT reveal
chain-of-thought.
- Do NOT mention model internals
(attention/scores/features/embeddings).
- Do NOT claim confirmed fraud; use cautious language
(suggests/likely/consistent with).
- Ground every claim in the given numbers/events.
- Use concrete numbers and specific timestamps/dates.
```

C.2 P_{Evidence}

```
 28-day ticketing window summary (up to final purchase).

INPUT
Numerical summary
- Purchase: Price T across 11
- Refund: Price R across 3

Key events (may be out of order; treat as evidence)

- [2024-09-14 10:26:00] (dep_date=2024-10-11) Station
A→Station C (Fri 12:00 departure) | REFUND | seats=1
| refunded Price A (fee Price F) | refunded 39,005
min before departure; held 68 min | usage-context:
same-route purchases in window: 3; unique routes in
window: 2 | anomaly-signals: adjacent-seat refund observed;
recent same-route refunds: 2; recent same-route refund
rate: 0.67; recent same-route refund amount: Price R
...
```

C.3 P_{Reason}

```
  Choose exactly ONE user type and ONE bulk-refund
pattern label.

User types:
- Commuter (weekday commute, fixed route/time)
- Weekend home-and-return (Fri outbound, Sun/Mon return)
- Weekend/leisure traveler (Fri–Sun, diverse routes/times)
- Weekend hospital visitor (Sat/Sun or fixed day, hospital
area routes)
- One-time user (short-term, little repetition)
- Corporate booking agent (books/cancels for others)

Bulk-refund patterns:
- Adjacent-seat holding + last-minute refunds
- Card-spend targeting via bulk purchase/refund
- Macro-assisted booking for resale
- Bulk seat-holding + selective refunds

TASK (do NOT output steps):
1) Pick one user type + one pattern label.
2) Justify BRB risk using the strongest event facts.
3) Explain deviations from the chosen type; infer ONE
plausible intent (cautious).
```

C.4 P_{Explain}

```
 OUTPUT :

=== NUMERICAL SUMMARY ===
- <purchase amount> across <purchase count>
- <refund amount> across <refund count>

=== BRB RISK JUSTIFICATION (<bulk-refund pattern label>)
===
4–6 sentences. Use concrete counts/amounts/timing; mention
last-minute/holding/adjacent seats if present.
```

Major Category	Minor Category	Feature Name	Description	Type	Mean ± Std
Ticketing Attributes	Identifier	dep_station_id	Departure station code	Categorical	-
		arr_station_id	Arrival station code	Categorical	-
		route_id	Encoded identifier of a departure–arrival station pair	Categorical	-
		train_id	Train service identifier	Categorical	-
	Transaction	action_type	Transaction type indicating booking (0) or refund (1)	Binary	0.29 ± 0.45
		seat_cnt	Number of seats involved in a single transaction	Numerical	1.00 ± 0.11
		buy_amt	Ticket purchase amount recorded at booking (KRW)	Numerical	19830.97 ± 19267.36
		refund_amt	Amount refunded at cancellation (KRW)	Numerical	8315.34 ± 16125.52
		cancel_fee	Total cancellation or refund fee charged (KRW)	Numerical	87.51 ± 546.75
		route_dist	Travel distance of the route (km)	Numerical	214.14 ± 115.13
	Time Info	travel_time	Scheduled travel duration (min)	Numerical	130.45 ± 1421.43
		lead_time_buy	Time interval between booking and departure (min)	Numerical	10655.97 ± 15146.40
		lead_time_ref	Time interval between refund and departure (min)	Numerical	1800.98 ± 6122.65
		hold_time	Duration between booking and refund timestamps (min)	Numerical	3039.76 ± 8913.92
		dep_dow	Day of week of departure	Categorical	-
		dep_hour	Hour of day of departure	Ordinal	-
Usage Intention	Forward History	route_buy_cnt	Cumulative number of bookings for the same route within the window	Numerical	5.02 ± 22.34
		fwd_dep_hour_median	Median departure hour of active forward tickets	Ordinal	-
		fwd_dep_dow_median	Median departure weekday of active forward tickets	Categorical	-
	Reverse Pattern	rev_buy_cnt	Cumulative number of bookings in the reverse direction	Numerical	4.13 ± 21.86
		rev_dep_hour_median	Median departure hour of active reverse tickets	Ordinal	-
		rev_dep_dow_median	Median departure weekday of active reverse tickets	Categorical	-
		rev_return_gap	Minimum time gap between arrival and reverse-direction departure (min)	Numerical	929.29 ± 3974.82
	Diversity	unique_route_cnt	Number of distinct routes used by the customer	Numerical	4.32 ± 15.06
Suspicious Signals	Active Pattern	overlap_cnt	Count of simultaneously held tickets with overlapping schedules	Numerical	0.54 ± 12.64
		same_route_cnt	Number of concurrently held tickets on the same route	Numerical	2.40 ± 17.59
		rev_route_cnt	Number of concurrently held tickets in the reverse direction	Numerical	2.17 ± 17.25
		repeat_interval	Median time interval between recent bookings on the same route (min)	Numerical	1403.81 ± 3556.46
		adj_seat_refund_flag	Indicator of whether adjacent seats were refunded together	Binary	0.04 ± 0.20
	Same Route Refund	recent_ref_cnt	Number of refunds for the same route within the window	Numerical	2.32 ± 6.51
		recent_ref_amt	Total refunded amount for the same route within the window (KRW)	Numerical	66828.88 ± 232610.97
		recent_ref_rate	Proportion of refunded bookings for the same route within the window	Numerical	0.30 ± 0.52

Table 4: Expert-guided sequential features for BRB detection.

```
=== BOOKING INTENT INTERPRETATION (<user type>) ===
3–5 sentences. Start: Selected user type: <...>. End:
A plausible intent is that ... (cautious).

=== DETAILED EVENT SUMMARY ===
Chronological (earliest→latest).
For EACH event:
- Describe the event factually (time, route, action,
amount).
- THEN select ONLY the most relevant 2–4 behavioral
signals that help interpret
(a) bulk-refund risk or (b) booking intent.
- Do NOT list all available features; include only those
that materially support interpretation.

Each line format (flexible, but concise):

   YYYY-MM-DD HH:MM:SS ACTION
For YYYY-MM-DD (Dow HH:MM) Origin→Destination | seats=<n>
| amount=<...> | key signals: <signal A>; <signal B>;
<signal C>
```