Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Weakly Supervised Detection of Operational Anomalies in Vaccine Refrigerators from Large-Scale Field Telemetry

**Abdoul Aziz Bonkoungou**

abdoul.bonkoungou@uni.lu

University of Luxembourg

**Navid Khaledian**

University of Luxembourg

**Abdoul Kader Kabore**

University of Luxembourg

**Ozgün Sakallı**

B Medical Systems

**Simon Justen**

B Medical Systems

**Josephine Alida Kabange**

B Medical Systems

**Kevin Laville**

B Medical Systems

**Jacques Klein**

University of Luxembourg

**Tegawendé F. Bissyandé**

University of Luxembourg

**Article**

**Keywords:** Vaccine Cold-chain monitoring, Door-opening events, Thermal response, Anomaly Detection, Weak Supervision, XGBoost

**Posted Date:** July 9th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9341655/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Weakly Supervised Detection of Operational Anomalies in Vaccine Refrigerators from Large-Scale Field Telemetry

Abdoul Aziz Bonkoungou[1*], Navid Khaledian[1], Abdoul Kader Kabore[1], Özgün Sakallı[2], Simon Justen[2], Josephine Alida Kabange[2], Kevin Laville[2], Jacques Klein[1], Tegawendé F. Bissyandé[1]

[1*]SnT, University of Luxembourg, 29, avenue John F. Kennedy, Luxembourg, 1855, Luxembourg.
[2]B Medical Systems, 17, Op der Hei, Hosingen, 9809, Luxembourg.

*Corresponding author(s). E-mail(s): abdoul.bonkoungou@uni.lu;
Contributing authors: navid.khaledian@uni.lu; abdoulkader.kabore@uni.lu; ozgun.sakalli@bmedicalsystems.com; simon.justen@bmedicalsystems.com; JosephineAlida.Kabange@bmedicalsystems.com; Kevin.Laville@bmedicalsystems.com; jacques.klein@uni.lu; tegawende.bissyande@uni.lu;

**Abstract**

Vaccine potency depends on an unbroken cold chain, yet the thermal stability of refrigerators in the field is shaped as much by human interaction patterns as by equipment design. Although door openings are an unavoidable part of daily vaccine management operations, no prior study has jointly quantified user behavior and the resulting thermal response at global scale using continuous, real-world telemetry. We address this gap with a large-scale empirical analysis of millions of timestamped observations collected from 500 vaccine refrigerators deployed across 32 countries. We propose a robust event-extraction method that converts raw sensor streams into semantically structured door episodes, each annotated with thermal rise, recovery time, burst frequency, and contextual features such as ambient temperature and operating-hour schedules. Building on this representation, we define six interpretable anomaly categories that span behavioral deviations (off-schedule access, excessive duration, bursty openings),

thermal deviations (thermal mismatch, slow recovery), and telemetry-quality issues. A cross-anomaly dependency analysis reveals that burst patterns and prolonged openings frequently compound into slow temperature recovery, while off-schedule usage acts as a persistent precursor to these thermal failures. Using the rule-derived labels as weak supervision, we train and evaluate three machine learning models for multi-label anomaly detection. XGBoost achieves an F1-score of 0.97 for binary anomaly detection and an average per-category F1-score of 0.82, outperforming Isolation Forest and Linear SVM across IID, country-held-out, and temporal splits. The learned model remains stable under up to 20% simulated label noise, confirming the viability of heuristic weak supervision for this task. These results establish a quantitative link between operational behavior and refrigerator thermal performance, providing a foundation for proactive, behavior-aware cold-chain monitoring in global immunization programs.

**Keywords:** Vaccine Cold-chain monitoring, Door-opening events, Thermal response, Anomaly Detection, Weak Supervision, XGBoost

# 1 Introduction

Vaccines are among the most cost-effective interventions in public health, yet their potency hinges on an unbroken cold chain from manufacturer to patient. Most vaccines are thermosensitive, and even brief deviations from the recommended storage range of +2 °C to +8 °C can degrade immunogenicity irreversibly [1–3]. Despite decades of investment in refrigeration technology and supply-chain infrastructure, temperature excursions remain widespread in low- and middle-income settings, where equipment failures and logistical constraints compound the challenge [4–6]. A growing body of evidence suggests that a significant share of these excursions originates not from hardware faults but from the way health workers interact with the equipment on a daily basis [7].

Door openings are the primary mechanism through which human behavior disturbs the thermal equilibrium of a vaccine refrigerator. Every time a door is opened for routine immunization, inventory checks, or vaccine retrieval, warm ambient air enters the cabinet and displaces the cold interior volume. The magnitude and duration of the resulting temperature perturbation depend on a combination of factors, including how long the door stays open, how frequently it is opened within a short window, and at what time of day the opening occurs [7, 8]. Prolonged openings, clustered "bursty" episodes, and access during non-standard hours can push internal temperatures toward or beyond the upper safety boundary, especially when the refrigerator is already under thermal stress from high ambient temperatures or a heavy vaccine load [9–11].

The thermal response of a refrigerator to a door opening, however, is not determined by user behavior alone. The rate at which the internal temperature rises and the speed of subsequent recovery are jointly influenced by environmental factors (ambient temperature, humidity), physical characteristics of the unit (insulation, compressor capacity, airflow design), and the stored payload [7, 12]. Performance also varies across refrigerator models due to differences in control logic, defrost cycles, and thermostat

calibration [12]. Studies in pharmaceutical and food refrigeration have confirmed these interdependencies between environmental conditions, usage intensity, and thermal stability [9–11]. Taken together, these findings highlight a complex interplay between human actions and device physics that cannot be captured by monitoring temperature alone.

Despite these observations, no large-scale study has jointly analyzed how users interact with cold-chain equipment and how those interactions translate into measurable thermal responses. Existing work falls into three largely disjoint categories. Temperature excursion surveys report the frequency and severity of out-of-range events at national or regional level but do not attribute excursions to specific operational causes [5, 6, 13]. Equipment audits assess compliance with storage guidelines at the point of inspection without continuous behavioral monitoring [1, 7]. Predictive modeling studies have begun to link door openings to thermal outcomes, but they rely on small-scale experimental or simulated data that do not reflect the heterogeneity of real-world deployments [8]. What remains absent is an integrated, data-driven framework that simultaneously captures user behavior and device thermal response across a large and geographically diverse fleet of vaccine refrigerators.

This study addresses that gap. Drawing on millions of timestamped telemetry records from 500 vaccine refrigerators deployed across 32 countries, we present an end-to-end framework for operational anomaly analysis that bridges the divide between behavioral monitoring and thermal assessment. The framework makes three main contributions.

- **Global-scale behavioral and thermal characterization.** We introduce a robust event-extraction method that transforms raw, irregularly sampled sensor streams into semantically structured door episodes, each annotated with duration, thermal rise, recovery time, burst context, and ambient conditions. Statistical analysis of these episodes reveals the circadian structure of access behavior, the nonlinear dependence of recovery time on thermal disturbance magnitude and door-open duration, and the compounding effect of bursty access on thermal load.
- **Interpretable operational anomaly taxonomy.** We define six anomaly categories (A1 through A6) spanning behavioral irregularities (off-schedule access, excessive duration, bursty openings), abnormal thermal responses (thermal mismatch, slow recovery), and telemetry-quality issues. Cross-anomaly analysis uncovers strong dependencies, most notably that off-schedule usage (A1) serves as a persistent precursor to slow temperature recovery (A5), and that burst patterns (A3) and long-duration openings (A2) frequently compound into thermal failures.
- **Weakly supervised multi-label anomaly detection.** Using the rule-derived anomaly labels as weak supervision, we train and evaluate three machine learning models under IID, country-held-out, and temporal data splits. XGBoost achieves an F1-score of 0.97 for binary anomaly detection and 0.82 on average across individual anomaly categories, while remaining stable under up to 20% simulated

label noise. A SHAP-based feature importance analysis confirms that device-relative z-score features and thermal context variables are the primary drivers of classification performance.

By explicitly linking patterns of human interaction to corresponding thermal outcomes at global scale, this study provides a quantitative foundation for transitioning cold-chain monitoring from reactive, threshold-based alarms toward proactive, behavior-aware operational intelligence.

## 2 Related Work

Research relevant to this study spans four areas: cold-chain temperature monitoring and compliance, the thermal effects of door-opening behavior in refrigeration systems, anomaly detection in IoT sensor networks, and weak supervision as a strategy for generating training labels without manual annotation. We review each area in turn and conclude by positioning the present work relative to the gaps that remain.

***Temperature monitoring and cold-chain compliance.***

Cold-chain reliability has been a longstanding concern in global immunization logistics. The dominant monitoring paradigm relies on electronic data loggers and fixed temperature thresholds to trigger alarms or generate retrospective compliance reports [4, 14, 15]. National and sub-national surveys have documented the frequency of temperature excursions across supply-chain tiers, revealing that exposure to both heat and freezing remains common even in settings with modern equipment [5, 6, 13]. Fahrni et al. [16] reviewed cold-chain logistics challenges that emerged during the COVID-19 vaccine rollout and emphasized the need for end-to-end temperature visibility. Robertson et al. [17] reviewed innovations in WHO-prequalified cold-chain equipment and emphasized that technological advances must be complemented by effective deployment and system-level integration to achieve their full impact. While these studies provide valuable epidemiological and logistical baselines, they treat temperature as an isolated signal and do not connect observed excursions to the specific user interactions or device conditions that caused them.

***Door-opening behavior and thermal response in refrigeration.***

A smaller body of work has examined the relationship between door openings and internal temperature dynamics. Chojnacky and Rodriguez [7] measured the effect of thermal ballast loading on domestic refrigerators used for vaccine storage and showed that the magnitude of temperature fluctuations depends heavily on payload volume and door-open duration. Martinez-Martinez et al. [9] studied how ambient temperature, food load, and door-opening frequency affect a preservation performance indicator in household refrigerators, demonstrating a nonlinear interaction among these factors. De Frias et al. [10] varied the frequency and duration of door openings in a commercial display case and reported corresponding effects on product temperature and energy consumption. In the vaccine domain specifically, Habiyaremye et

al. [8] trained a multiple linear regression model on data from a thermoelectric vaccine refrigerator to predict time-to-threshold after a door opening. Although these studies confirm that user behavior shapes thermal outcomes, they are limited to controlled laboratory environments or small experimental datasets and do not scale to heterogeneous, multi-country field deployments.

***Anomaly detection in cold-chain and IoT sensor networks.***

Anomaly detection within vaccine cold-chain monitoring remains a nascent area. Putrada and Abdurohman [18] compared histogram-based outlier scoring (HBOS), minimum covariance determinant (MCD), and one-class SVM on 11 days of one-minute temperature readings from a single vaccine refrigerator, a dataset containing only 89 outliers among 17,072 normal observations. Harrabi et al. [19] trained a semi-supervised convolutional autoencoder to reconstruct normal temperature profiles from 24-hour recording windows and flagged reconstruction errors as anomalies. More recently, Harrabi et al. [**?** ] deployed a lightweight deep-learning model on a resource-constrained microcontroller for real-time vaccine temperature anomaly detection, demonstrating the feasibility of edge inference but again operating on a single-device, short-duration dataset. All three studies focus exclusively on temperature readings and do not incorporate behavioral signals such as door events.

The broader literature on time-series and IoT anomaly detection offers a rich methodological toolkit. Boniol et al. [20] provide a decade-spanning review of time-series anomaly detection that covers forecasting-based, reconstruction-based, and representation-learning approaches. Zamanzadeh Darban et al. [21] survey deep learning methods for the same task, including LSTM autoencoders, temporal convolutional networks, and transformer-based architectures. IoT-focused reviews emphasize practical challenges that are highly relevant to cold-chain telemetry, including irregular sampling intervals, environmental noise, missing data, and severe class imbalance [22, 23]. Hybrid supervised-unsupervised pipelines have shown promise in industrial multi-sensor settings [24, 25], but these approaches typically require large volumes of accurate labels and offer limited interpretability, constraints that are difficult to satisfy in global-health field deployments.

***Weak supervision and programmatic labeling.***

When reliable ground-truth annotations are scarce or expensive to obtain, weak supervision offers a practical alternative by using heuristic rules, knowledge bases, or noisy labeling functions to generate training signal at scale. The data-programming paradigm introduced by Ratner et al. [**?** ] formalized this idea in the Snorkel framework, where multiple noisy labeling functions are composed through a generative model to produce probabilistic labels. Subsequent work extended the concept to multi-task settings [**?** ] and demonstrated its viability in clinical domains where manual annotation is particularly costly [**?** ]. Jiang et al. [26] surveyed weakly supervised anomaly detection more broadly and identified rule-based label generation as a promising direction for operational monitoring contexts where labeled anomalies are rare by definition.

Our work adopts this paradigm by encoding domain knowledge about vaccine refrigerator operation into six interpretable labeling rules and using the resulting labels to train a supervised multi-label classifier.

***Positioning of this work.***

The present study bridges several of these threads. Unlike prior cold-chain studies that monitor temperature in isolation, we jointly model user behavior and thermal response. Unlike laboratory studies of door-opening effects, we operate at the scale of 500 devices across 32 countries using continuous field telemetry. Unlike general-purpose time-series anomaly detectors, we embed domain-specific operational semantics into the anomaly taxonomy. And unlike approaches that require clean labeled data, we derive weak supervision from interpretable heuristic rules and demonstrate that a gradient-boosted classifier trained on these labels generalizes across geographies, time periods, and label noise levels.

# 3 Methodology

This section describes the dataset, the overall framework, the event-construction procedures used to transform raw telemetry into structured door episodes, and the six anomaly categories that define the labeling taxonomy. We first present the data and the two-stage framework overview (labeling and detection), then detail the algorithms for episode extraction, battery-event identification, and operating-hour inference, and finally specify each anomaly rule.

## 3.1 Dataset

The dataset used in this study comprises operational time-series measurements collected from 500 representative TCW 40 SDD vaccine refrigerators manufactured by a leading company specializing in vaccine cold-chain equipment. These units are deployed across 32 countries in Africa, South America, South Asia, and other regions. Each unit is equipped with a data-logging device that captures the refrigerator's operational state under real-world field conditions.

The resulting dataset is a multivariate time series containing four primary sensor channels: internal temperature, ambient temperature, door status (open or closed), and logger battery level. Baseline measurements are recorded at a regular ten-minute sampling interval to track steady-state behavior. Certain high-priority events, such as the exact moment a door is opened or a battery fluctuation, are captured asynchronously as they occur. The dataset therefore combines synchronous, periodically sampled data with asynchronous, event-driven signals. This hybrid structure is essential for accurately representing both the background environmental conditions and the precise timing of user interactions or device state changes.

## 3.2 Framework overview

### 3.2.1 Anomaly labeling

Figure 1 illustrates the anomaly labeling strategy. Continuous sensor streams, including door state, internal temperature, and logger battery level, are first processed by an event detector that segments the data into semantically meaningful door-opening episodes, each corresponding to a complete open-to-close interaction. From these door events, we infer per-device operating hours, analyze the burstiness and duration of openings, and extract thermal response characteristics. These features, together with battery-derived events, are then evaluated through six interpretable rule families (A1 through A6) to identify behavioral, thermal, and operational anomalies. Each rule captures a distinct deviation mode, and multiple rules can fire for the same event, producing a composite anomaly label. Details on the anomaly types and their labeling logic are given in Sections 3.3 and 3.4.

This labeling framework serves two complementary purposes. First, it provides an immediate, explainable diagnostic layer that field engineers can use for automated quality assurance and domain-specific root-cause analysis. Second, it generates structured weak labels that can be used to train or validate machine learning models for automatic anomaly detection at scale.

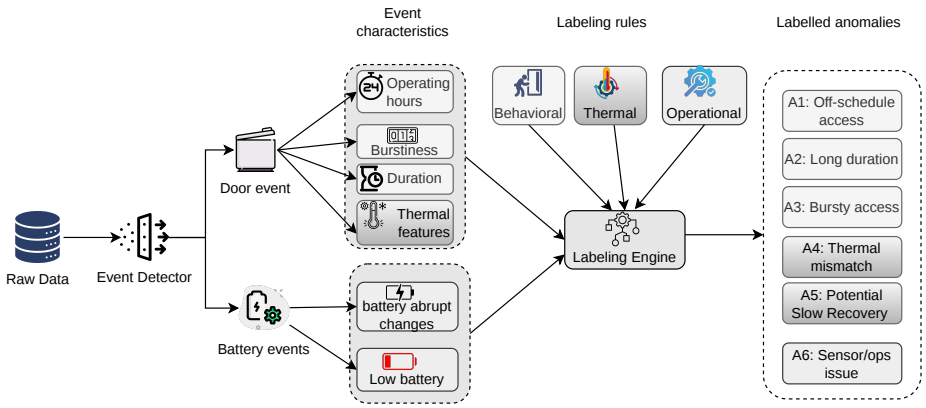



**Fig. 1 Overview of the anomaly labeling framework.** Raw door, temperature, and battery data are transformed into door episodes, per-device operating-hour masks, and battery-derived signals. Six rule families (A1 through A6) label each episode. Any active rule marks the event as anomalous and records the corresponding reasons.

### 3.2.2 Anomaly detection pipeline

Figure 2 illustrates the end-to-end pipeline for detecting anomalous refrigeration behavior from the labeled data. Starting from the episodes labeled via the heuristic rules (A1 through A6), we extract three complementary groups of features.

Episode-level features characterize event dynamics, including duration, peak internal temperature rise $\Delta T_{\text{peak}}$, time-to-peak, and recovery time. Contextual features capture operating conditions such as time of day, day of week, country, and ambient temperature statistics. Device-history features encode deviations relative to each device's own historical behavior through metrics such as burst frequency and z-scored temperature, recovery, duration, and burst-intensity values that provide longitudinal context.

These engineered features serve as inputs for training and validating machine-learning-based anomaly detectors under weak supervision. To rigorously test model generalizability and robustness to distribution shifts, we employ three splitting strategies. An IID split provides baseline performance. A country split evaluates adaptation to previously unseen geographic deployments. A time split assesses stability across future temporal horizons. Three model families (XGBoost, Isolation Forest, and Linear SVM) are trained to classify anomalies, with performance evaluated using F1-score and Average Precision (AP) on both binary (any anomaly present) and multi-label (per-category) tasks.

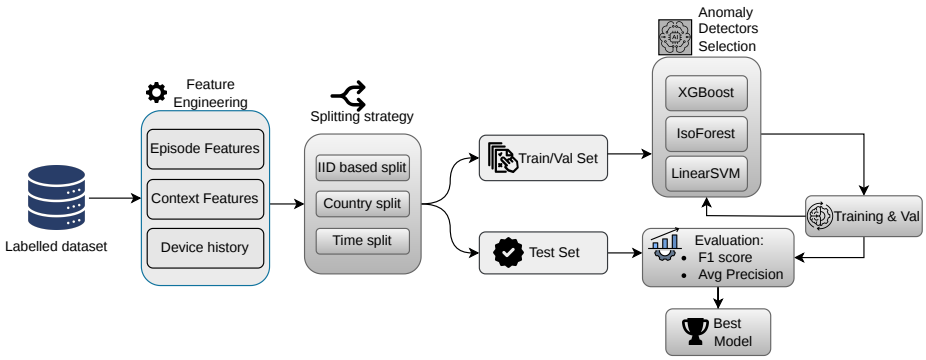



**Fig. 2 End-to-end anomaly detection pipeline.** Heuristic-labeled episodes are transformed into three feature groups and then fed to supervised and unsupervised classifiers. Training and testing follow three data-split strategies to measure generalization, with F1 and AP computed for binary and per-category anomaly scores.

## 3.3 Event construction

### 3.3.1 Door episode extraction

Raw door-state data are segmented into *door episodes*, defined as contiguous open-to-close intervals for each device. For each episode, we compute its duration and a set of thermal response features derived from the internal temperature signal. These features include a pre-opening baseline, the maximum post-opening temperature rise ($\Delta T_{\text{peak}}$), the time required to reach that peak, and a recovery time measuring how quickly the temperature returns toward the baseline after the peak. Recovery is defined as the first point in time at which the internal temperature falls below a threshold set

as a fraction $\alpha$ of the peak rise above the baseline, within a bounded search horizon. Ambient temperature is summarized alongside these features through its mean before opening, its value at the moment of the internal temperature peak, and its mean during the recovery search window. Algorithm A details the full procedure for extracting door episodes and computing their associated thermal features. These episode-level features align user interactions with the refrigerator's thermal response, providing interpretable behavioral and physical indicators for downstream anomaly detection.

**Algo A: Door Episode Extraction with Thermal Features**

**Inputs:**

- Per-device stream $\{(t, s_t, T_t^{\text{int}}, T_t^{\text{amb}})\}$ with door state $s_t \in \{0, 1\}$.
- Baseline window $W_b$ minutes; post-open peak horizon $W_p$ minutes.
- Recovery horizon $W_r$ minutes; recovery fraction $\alpha \in (0, 1]$.

**Algorithm:**

1. For each device $d$, sort by time.
2. Identify door episodes (open-to-close intervals). If the stream ends while open, close the episode at the last observed open timestamp.
3. For each episode with start $t_{\text{open}}$ and end $t_{\text{close}}$, compute:
   (a) Duration $D = (t_{\text{close}} - t_{\text{open}})$.
   (b) Baseline $B = \text{mean}\{T_t^{\text{int}} : t \in [t_{\text{open}} - W_b, t_{\text{open}})\}$ and ambient baseline `ext_baseline` $= \text{mean}\{T_t^{\text{amb}} : t \in [t_{\text{open}} - W_b, t_{\text{open}})\}$.
   (c) Peak within $[t_{\text{open}}, t_{\text{open}} + W_p]$: $\Delta T_{\text{peak}} = \max(T^{\text{int}}) - B$, time-to-peak, and ambient at peak `ext_at_peak` $= T_{t_{\text{peak}}}^{\text{amb}}$.
   (d) Recovery search within $[t_{\text{peak}}, t_{\text{peak}} + W_r]$: first $t$ with $T_t^{\text{int}} \leq B + \alpha\, \Delta T_{\text{peak}}$; report minutes since peak as `recovery_min`. Also compute mean ambient within the search window `ext_mean_recovery`.

**Output:** One row per episode with {`device_id, start, end, duration_sec, dT_peak, t_to_peak_min, recovery_min, ext_baseline, ext_at_peak, ext_mean_recovery`}.

### 3.3.2 Battery change events and low-battery episodes

We derive two operational signals from the battery telemetry. The first is a set of discrete battery change events, defined as drops or rises exceeding a minimum step. The second is a set of contiguous low-battery episodes during which the battery state remains below a fixed threshold. Battery change events are extracted by computing successive differences and retaining steps with $|\Delta B| \geq \varepsilon$, while debouncing events that occur within a short temporal gap. Low-battery episodes are identified as runs with $B_t < \theta$, which are split when the gap between consecutive low readings exceeds a break threshold and discarded if their duration is shorter than a minimum requirement. Algorithm B summarizes the extraction logic.

Variations in battery level may occasionally affect telemetry quality, leading to intermittent uploads, delayed reporting, or transient sensor artifacts. Explicitly identifying these periods improves the reliability of anomaly attribution by helping distinguish genuine behavioral or thermal deviations from irregularities caused by logger power instability.

**Algo B: Battery Events and Low-Battery Episode Extraction**

**Inputs:**

- Per-device battery series $\{(t, B_t)\}$ with $B_t \in [0, 100]$ (%).
- Event step threshold $\varepsilon$; debounce gap $G_{\min}$.
- Low threshold $\theta$; minimum low duration $L_{\min}$; break gap $G_{\text{break}}$.

**Algorithm (Change Events):**

1. For each device, sort by time; compute $\Delta B_t = B_t - B_{t^-}$.
2. Keep times with $|\Delta B_t| \geq \varepsilon$; ignore those within $G_{\min}$ minutes of the previous kept event.
3. Emit $\{t, \Delta B_t, B_{t^-}, B_t, \text{direction}, \Delta t_{\text{prev}}\}$.

**Algorithm (Low Episodes):**

1. Mark indices with $B_t < \theta$; split the run whenever time between consecutive low readings exceeds $G_{\text{break}}$ minutes.
2. For each run, compute start/end, duration, and $\min B_t$; keep only if duration $\geq L_{\min}$.

**Output:**

- *Battery change events:* `[device_id, timestamp, delta_battery, prev_battery, new_battery, direction, dt_prev_event_min]`.
- *Low-battery episodes:* `[device_id, low_start, low_end, duration_min, min_battery]`.

### 3.3.3 Inferring operating hours from door events

We infer the typical operating hours of each device, that is, the hours of the day when door activity consistently occurs, by analyzing the temporal distribution of historical door-opening events. This yields a per-device hourly mask $M_{d,h}$ identifying expected periods of use.

For each device, door-open timestamps are grouped by day and hour to compute an average hourly open rate. The resulting 24-hour profile is circularly smoothed with a three-hour window to reduce noise and emphasize consistent activity patterns. We then rank hours by their smoothed activity and select the smallest subset that accounts for 90% ($\alpha = 0.90$) of total daily activity, with bounds enforcing between 1 and 18 hours per day. This data-driven approach captures when each device is typically accessed, accounting for device-specific usage patterns that may vary across clinics, regions, and operational contexts.

The inferred schedule serves as a behavioral baseline that distinguishes expected operational periods (e.g., daytime clinic hours) from off-schedule events. It enables contextual anomaly detection and supports fair comparison across devices with differing usage rhythms.

**Algo C: Operating-Hours Inference from Door Events**

**Goal:** Infer per-device "operating hours" (hours-of-day with consistent human activity) from door-open events.

**Inputs:**

- Door-open timestamps $\{(d_i, t_i)\}$ with device IDs $d_i$
- Minimum active days $D_{\min}$ (default: 5)
- Coverage fraction $\alpha \in (0, 1]$ (default: 0.90)
- Circular smoothing window size $2w{+}1$ hours (default: $w = 1$, giving 3-hour window)
- Bounds on selected hours: $H_{\min} = 1, H_{\max} = 18$

**Algorithm:**

1. For each device $d$, collect door-open timestamps $\{t_i\}$.
2. Map each $t_i$ to its day $u_i = \lfloor t_i \rfloor_{\text{day}}$ and hour $h_i \in \{0, \ldots, 23\}$.
3. Compute daily hourly counts:

$$C_{d,u,h} = \#\{\, i : (u_i, h_i) = (u, h) \,\}.$$

4. Let $N_d$ be the number of distinct *active* days for device $d$ (days with $\sum_h C_{d,u,h} > 0$).
   Discard devices with $N_d < D_{\min}$.
5. Compute mean hourly open rate across active days:

$$R_{d,h} = \frac{1}{N_d} \sum_u C_{d,u,h}, \quad h = 0, \ldots, 23.$$

6. Apply circular smoothing over the 24-hour cycle:

$$S_{d,h} = \frac{1}{2w+1} \sum_{k=-w}^{w} R_{d,(h+k) \bmod 24}.$$

7. Let $T_d = \sum_{h=0}^{23} S_{d,h}$. Order hours by descending $S_{d,h}$ and select the smallest set:

$$H_d^\star = \arg \min_{H \subseteq \{0,\ldots,23\}} \Big\{ |H| : \sum_{h \in H} S_{d,h} \geq \alpha\, T_d \Big\}.$$

8. Enforce bounds:
   If $|H_d^\star| < H_{\min}$, add next-best hours until $|H_d^\star| = H_{\min}$.
   If $|H_d^\star| > H_{\max}$, keep top $H_{\max}$ hours by $S_{d,h}$.
9. Define hourly mask:

$$M_{d,h} = \mathbb{1}[h \in H_d^\star], \quad h = 0, \ldots, 23.$$

**Output:** For each device $d$, an hourly mask $M_{d,h} \in \{0, 1\}$ indicating inferred operating hours

Algorithm C details the inference of per-device operating hours from historical door-opening patterns. By aggregating and smoothing hourly activity, it identifies consistent periods of human interaction that characterize the device's regular schedule. These inferred masks later serve as behavioral baselines for detecting off-schedule or contextually abnormal events.

## 3.4 Anomaly types

We define six complementary anomaly categories (A1 through A6) that capture distinct modes of deviation in refrigerator operation and data integrity. These rules integrate door-event statistics, thermal response features, and device telemetry to identify unusual behavior. The resulting framework unifies behavioral, thermal, and operational indicators in a consistent per-device labeling pipeline. Table 1 summarizes the detection logic and key parameters for each category.

**Table 1** Summary of anomaly rules (A1 through A6) combining behavioral, thermal, and operational indicators.

| Code | Name | Detection logic | Key parameters |
|---|---|---|---|
| A1 | Off-schedule use | Hour outside inferred open-hour mask $\mathsf{M}_{d,h} = 0$. | $\alpha = 0.90$ (coverage) |
| A2 | Long duration | duration $>$ $\mathrm{p95}_{d,h}$ or $> Q3_d + 1.5\,\mathrm{IQR}_d$. | $p = 0.95$, $k_{\mathrm{iqr}} = 1.5$ |
| A3 | Bursty access | Count in 30 min window exceeds $\mu_d + 3\sigma_d$. | $w = 30$ min, $z = 3$ |
| A4 | Thermal mismatch | $\Delta T_{\mathrm{peak}} < 0.1^{\circ}$C (when not cold) or z-score$> 3$ (when not hot). | $\Delta T_{\mathrm{min}} = 0.1^{\circ}$C, $z = 3$ |
| A5 | Potential slow recovery | recovery_min $>$ $p95_d$ (multi-factor adjusted) or missing. | $q = 0.95$, $\beta = 0.2$ |
| A6 | Sensor/ops issue | Event overlaps low battery episode or battery drop $\Delta B \leq -10\%$ within $\pm 15$ min. | $\Delta B_{\mathrm{max}} = -10\%$, $t_{\mathrm{prox}} = 15$ min |

### 3.4.1 A1: Off-schedule usage

An event is classified as off-schedule when it occurs outside the refrigerator's inferred daily operating hours. Each device's typical access hours are derived from historical door activity using the smoothed 24-hour profile described in Algorithm C. For each event, we check whether its start hour falls outside the device's active-hours mask ($\mathsf{M}_{d,h} = 0$). Accessing a refrigerator during non-operating periods, such as late at night or on weekends, can indicate unauthorized use, procedural deviations, or emergency interventions that warrant further investigation.

### 3.4.2 A2: Excessive door-open duration

This category captures door episodes that last unusually long relative to typical patterns for the same device. Two complementary thresholds are applied. The first is the 95th percentile of event duration for the same hour of day ($\mathrm{p95}_{d,h}$). The second is a device-wide outlier rule based on the interquartile range, $Q3_d + 1.5 \times \mathrm{IQR}_d$. An event

exceeding either limit is flagged. Prolonged door openings may expose stored vaccines to unsafe temperatures, which can result from restocking procedures, equipment inspections, or malfunctioning door seals.

### 3.4.3 A3: Bursty access behavior

This category identifies clusters of many door openings occurring within a short time window. For each event, the algorithm counts how many openings occur within a sliding $\pm$30-minute window. A burst is flagged if this count exceeds $\mu_d + 3\sigma_d$, where $\mu_d$ and $\sigma_d$ are the device's mean and standard deviation of window counts. Repeated access over short periods may reflect abnormal usage such as stocktaking, sensor chattering, or operational disruptions.

### 3.4.4 A4: Thermal mismatch

This category flags temperature responses that are inconsistent with typical thermal dynamics following a door opening. Two complementary rules are applied with ambient-aware conditions. A low temperature rise is flagged when $\Delta T_{\text{peak}} < 0.1°\text{C}$ and the ambient temperature is not cold (above the device's 20th percentile), since minimal air exchange under warm ambient conditions is unexpected and may indicate sensor noise or a recording artifact. A high thermal excursion is flagged when the z-score $(\Delta T_{\text{peak}} - \mu_d)/\sigma_d > 3$ and the ambient temperature is not hot (below the device's 80th percentile), since extreme rises under moderate ambient conditions may indicate a door left open or a temperature sensor malfunction. The per-device statistics $\mu_d$ and $\sigma_d$ are estimated over all valid events for the device. The ambient-aware logic reduces false positives while maintaining sensitivity to true thermal anomalies.

### 3.4.5 A5: Potential slow recovery

This category identifies events where the refrigerator takes unusually long to return to its baseline temperature after a door opening. Each device's recovery time distribution is summarized by its 95th percentile. The threshold is dynamically adjusted based on five contextual factors that legitimately affect recovery time: ambient temperature during recovery, door-opening duration, temperature rise magnitude ($\Delta T_{\text{peak}}$), baseline temperature before opening, and recent burst activity. Each factor is normalized to $[0, 1]$ by scaling relative to its per-device median and 90th percentile, then weighted equally ($\beta = 0.2$) to compute a combined adjustment. The adjusted threshold equals $p95_d \times (1 + \sum_i \beta_i \cdot \text{factor}_i)$, allowing up to 100% increase under extreme conditions where all factors reach their 90th-percentile values. Events exceeding this adjusted threshold are flagged, as are those with missing recovery estimates. This multi-factor adjustment reduces false positives from legitimate long recoveries under challenging conditions (e.g., high ambient temperature, long door opening, large thermal load) while maintaining sensitivity to genuine cooling system anomalies.

### 3.4.6 A6: Sensor or operational anomalies

This category flags events occurring under conditions of power or battery instability. Two temporal conditions are evaluated. The first checks whether an event overlaps a

recorded low-battery episode ($[t_{\text{low}}^{(s)}, t_{\text{low}}^{(e)}]$). The second checks whether a large battery drop ($\Delta B \leq -10\%$) occurs within $\pm 15$ minutes of the event start. Matching is performed per device using interval overlap for episodes and nearest-neighbor temporal joins for battery drops. Low power can cause logging interruptions, delayed telemetry, or incorrect temperature sensing, and flagging such events helps isolate data artifacts from genuine behavioral anomalies.

Together, these six heuristic rules provide complementary coverage. Categories A1 through A3 capture behavioral irregularities in access timing or frequency. Categories A4 and A5 capture abnormal thermal responses of the refrigeration cycle. Category A6 captures telemetry or sensor-related reliability issues. Table 1 provides a consolidated view of the detection logic and parameter settings.

# 4 Results and Discussion

## 4.1 Statistical analysis of door-opening behavior and thermal dynamics

This subsection characterizes the empirical distributions of access behavior and thermal response across the full dataset. We examine door-opening durations, circadian access structure, thermal recovery dynamics, cross-anomaly interactions, and the overall prevalence of each anomaly category.

### 4.1.1 Door-opening duration distributions

Door-opening events exhibit a heavily right-skewed duration distribution (Figure 3). The median duration is 0.2 minutes and the 95th percentile is 3.3 minutes, confirming that the vast majority of openings are brief. The empirical cumulative distribution function (ECDF) shows that over 90% of episodes conclude within the first few minutes of exposure. A small number of extreme outliers, however, extend into much longer periods. Although rare, these prolonged episodes impose disproportionately high thermal load on the refrigerator interior and represent meaningful tail-risk events from the perspective of vaccine safety.

### 4.1.2 Circadian access patterns and off-schedule anomalies

Access activity follows a pronounced circadian rhythm (Figure 4). Normalized access frequency is lowest overnight, rises through the morning, and peaks in the late morning before declining through the afternoon. Off-schedule anomalies (A1), by contrast, concentrate during the nighttime and early-morning hours between approximately 00:00 and 03:00, when legitimate access is minimal. This inverse relationship between normal activity and off-schedule anomaly rate confirms that deviations from expected behavior are most detectable during low-activity periods and supports the value of time-of-day conditioning in anomaly detection models.

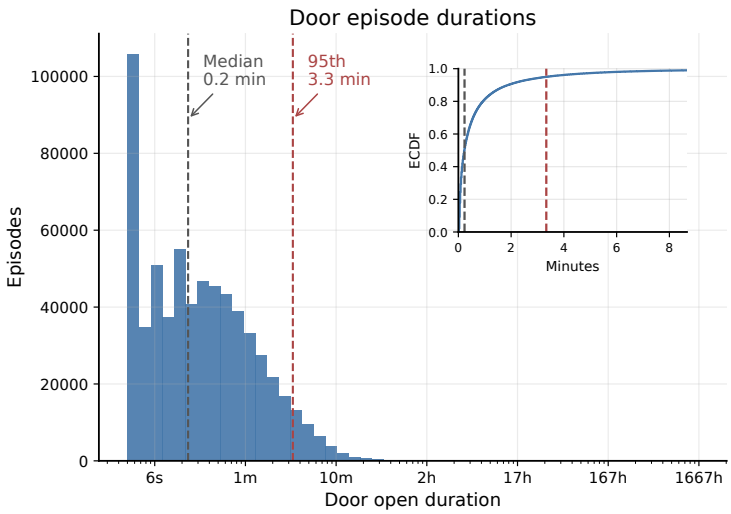



**Fig. 3** Distribution of door episode durations across the full dataset. The right-skewed distribution confirms that most openings are brief, with a median of 0.2 min and 95th percentile of 3.3 min.

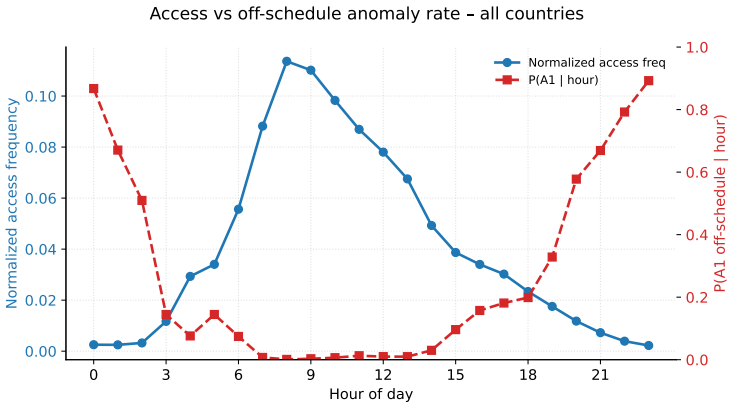



**Fig. 4** Normalized access frequency (left axis) and off-schedule anomaly rate (right axis) by hour of day. The two curves are approximately anti-correlated, with A1 anomalies concentrating during nighttime hours.

### 4.1.3 Thermal recovery dynamics

***Dependence on peak internal temperature rise.***

Recovery time increases nonlinearly with the peak internal temperature rise $\Delta T_{\text{peak}}$ (Figure 5). For small disturbances ($\Delta T_{\text{peak}} < 0.4°\text{C}$), median recovery time remains low and stable. Beyond this range, median recovery time rises sharply and the interquartile range widens considerably, indicating that larger thermal disturbances are associated with both slower and more variable recovery trajectories. This pattern

is consistent with the thermodynamic expectation that greater thermal displacement requires more compressor work and longer rebalancing time.

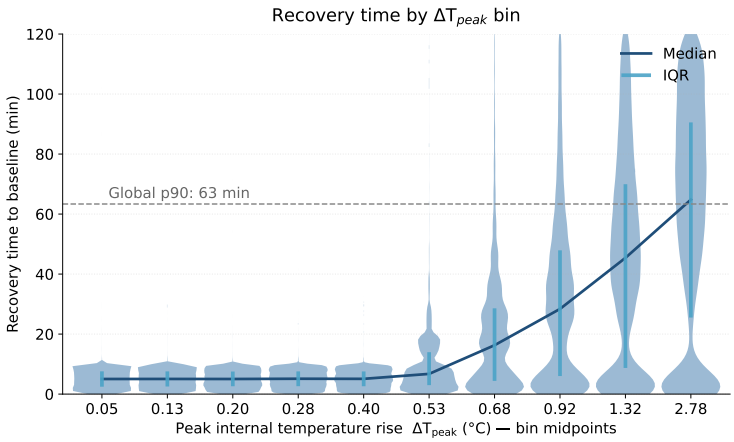



**Fig. 5** Recovery time as a function of peak internal temperature rise $\Delta T_{\text{peak}}$. Median recovery time increases nonlinearly beyond approximately 0.4°C, with widening variability at higher disturbances.

***Dependence on door-opening duration.***

A similar trend emerges when conditioning on door-open duration (Figure 6). Short openings of approximately one minute or less recover quickly, with median recovery times on the order of only a few minutes. Long openings of ten minutes or more, on the other hand, are associated with substantially longer recovery periods, with median times commonly in the 50 to 80 minute range and high variability. The global 90th-percentile recovery time across all episodes is 59 minutes, providing a practical upper bound for thermal re-stabilization under typical field conditions.

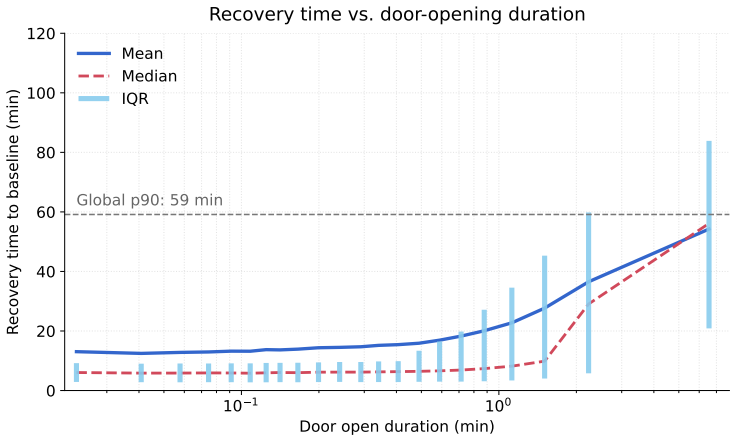



**Fig. 6** Recovery time as a function of door-opening duration. Prolonged openings (≥10 min) are associated with median recovery times in the 50–80 min range and high variability.

***Impact of bursty access.***

Episodes that co-occur with bursty access (A3) exhibit noticeably slower recovery than non-burst episodes (Figure 7). Compared to non-burst periods, A3 bursts are associated with substantially longer average recovery times. This pattern suggests that thermal load accumulates across a burst window even when individual door-open events are brief, because the compressor does not have sufficient time to return the cabinet to baseline between successive openings. The finding underscores the importance of burst-aware thermal models for accurate anomaly characterization.

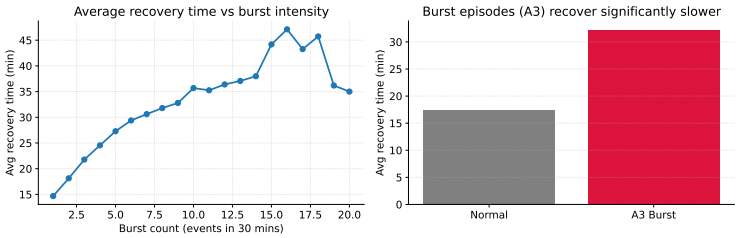



**Fig. 7** Comparison of recovery time distributions between burst (A3) and non-burst episodes. Bursty access is associated with substantially longer recovery periods.

### 4.1.4 Cross-anomaly interactions between A3 and A4

We observe a clear interaction between bursty access (A3) and thermal anomalies (A4). As shown in Figure 8, the probability of a high-$\Delta T$ A4 anomaly is substantially higher during A3 burst events than during non-burst periods, whereas low-$\Delta T$ anomalies are more prevalent outside burst windows. These conditional patterns indicate a strong directional association: repeated rapid openings drive elevated thermal disturbance,

while the absence of burst activity is more commonly associated with the opposite anomaly mode (minimal or absent temperature response to a door opening).

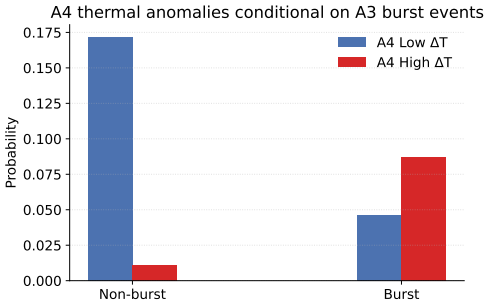



**Fig. 8** Conditional probability of A4 thermal anomaly subtypes given the presence or absence of bursty access (A3). High-$\Delta T$ anomalies concentrate during burst periods.

### 4.1.5 Global anomaly distribution

Figure 9 summarizes the empirical distribution of anomaly types. Potential slow recovery (A5) and low-$\Delta T$ thermal mismatch (A4) are the two most prevalent categories, each accounting for roughly 17% of all flagged anomalies. Long-duration events (A2) are the next most common category. By contrast, bursty access (A3) and high-$\Delta T$ thermal mismatch are relatively infrequent, while sensor or operational anomalies (A6) are negligible and will be excluded from the machine learning experiments in the remainder of this study. This prevalence pattern indicates that slow thermal recovery and moderate temperature deviations constitute the dominant departures from nominal operation across the global fleet, which is consistent with the thermodynamic analysis above showing that even short openings can compound into recovery challenges under unfavorable ambient or burst conditions.

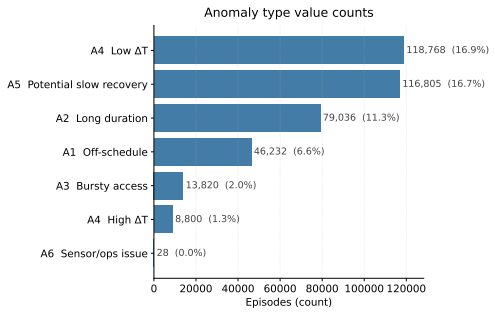



**Fig. 9** Distribution of anomaly types within the labeled dataset. A5 (slow recovery) and A4 (low $\Delta T$) are the most prevalent categories. A6 is negligible and excluded from ML experiments.

## 4.2 Weak supervision for anomaly detection

We evaluate the effectiveness of weak supervision for anomaly detection in three experiments addressing model comparison, feature importance, and robustness to label noise.

### 4.2.1 Experiment 1: Baseline model comparison

***Objective.***

Establish baseline performance by comparing supervised, unsupervised, and rule-based approaches across different data distributions, evaluating both detection accuracy and generalization.

***Experimental setup.***

We use 14 features organized into three groups. Episode-level features (6 features) describe the intrinsic characteristics and temporal context of each door episode: hour of day, day of week, door-opening duration, peak temperature deviation, time to peak, and recovery time. Thermal context features (3 features) capture environmental conditions influencing thermal dynamics: external baseline temperature, external temperature at the peak, and mean external temperature during recovery. Device-relative features (5 features) encode deviations relative to each device's historical behavior: burst count within a 30-minute window, z-scored temperature deviation, z-scored recovery time, z-scored duration, and z-scored burst intensity.

We evaluate three model families. XGBoost [27] is a supervised multi-label gradient-boosted tree ensemble. Isolation Forest [28] is an unsupervised anomaly detector that isolates anomalies through random recursive partitioning. Linear SVM [29] is a supervised linear classifier. We also retain the rule-based weak labels as a reference baseline.

To assess generalization, we employ three splitting strategies. The IID split uses a random 70/15/15 train/test/validation partition stratified by country. The country split trains on 80% of countries and tests on the remaining 20%. The time split trains on the first 70% of events chronologically and tests on the last 30%.

For XGBoost, we perform 40 iterations of random search with three-fold stratified cross-validation, optimizing average precision. For Isolation Forest, we set contamination to 0.06 based on the empirical anomaly prevalence. For Linear SVM, we tune the regularization parameter $C$ and class weights over 20 iterations of random search. All models are evaluated on held-out test sets using F1-score and average precision (AP). We report two complementary metric sets. The "ANY" metrics (F1, AP) measure binary anomaly detection, that is, whether any anomaly is present in an episode regardless of type. The "Average" metrics (Avg F1, Avg AP) measure multi-label classification performance by averaging per-label scores across all anomaly categories (A1 through A5), capturing the model's ability to distinguish specific anomaly types. F1-score balances precision and recall, while AP summarizes the precision-recall curve and is robust to class imbalance [30, 31].

**Table 2** Baseline model comparison across three data-splitting strategies.

| Split | Model | ANY F1 | ANY AP | Avg F1 | Avg AP |
|---|---|---|---|---|---|
| IID | Isolation Forest | 0.628 | 0.599 | – | – |
| | Linear SVM | 0.829 | 0.916 | 0.689 | 0.792 |
| | XGBoost | **0.876** | **0.953** | **0.779** | **0.879** |
| Country | Isolation Forest | 0.622 | 0.594 | – | – |
| | Linear SVM | 0.808 | 0.904 | 0.680 | 0.775 |
| | XGBoost | **0.839** | **0.931** | **0.798** | **0.854** |
| Time | Isolation Forest | 0.630 | 0.599 | – | – |
| | Linear SVM | 0.832 | 0.915 | 0.692 | 0.778 |
| | XGBoost | **0.924** | **0.982** | **0.878** | **0.930** |

– indicates unsupervised methods produce only binary predictions (no per-label metrics).

***Results and analysis.***

Table 2 presents the baseline comparison. XGBoost consistently achieves the strongest performance across all three splits. Binary detection F1 (ANY F1) ranges from 0.839 under the country split to 0.924 under the time split, with corresponding AP scores between 0.931 and 0.982. Per-label performance is also highest for XGBoost, with average F1 between 0.779 and 0.878 and average AP between 0.854 and 0.930.

Linear SVM provides competitive binary detection (ANY F1 between 0.808 and 0.832, ANY AP between 0.904 and 0.916) with lower model complexity, although its per-label classification consistently falls below XGBoost. Isolation Forest performs substantially worse in binary detection (ANY F1 between 0.622 and 0.630, ANY AP between 0.594 and 0.599), and as an unsupervised method it produces only binary predictions without per-label discrimination.

Generalization across splits is encouraging. Performance remains stable between the IID and country splits, with less than a four-percentage-point F1 difference for XGBoost, indicating reasonable cross-country robustness. The time split yields the highest performance (ANY F1 of 0.924), suggesting that temporal generalization is strong when training precedes testing chronologically, likely because the model benefits from learning the full range of seasonal and operational variation present in the earlier portion of the data.

### 4.2.2 Experiment 2: Feature ablation study

***Objective.***

Quantify the contribution of each feature group to model performance and identify the minimal feature set required for effective anomaly detection.

***Methodology.***

We evaluate XGBoost under four progressive feature configurations. The *Episode-only* configuration uses six basic episode characteristics: time of day, day of week, duration,

temperature rise, time to peak, and recovery time. The *Episode + Context* configuration adds three thermal context features: external baseline temperature, external temperature at peak, and mean external temperature during recovery (nine features total). The *Episode + Device History* configuration replaces the context features with five device-relative historical features: burst count within a 30-minute window, z-scored temperature deviation, z-scored recovery time, z-scored duration, and z-scored burst intensity (11 features total). The *Full* configuration combines all 14 features. For fair comparison, we use the same data split and hyperparameters as the best-performing configuration from Experiment 1.

***Results and analysis.***

Figure 10 presents the progressive impact of feature groups. The episode-only baseline achieves solid binary anomaly detection (ANY F1 of 0.863, ANY AP of 0.945) but struggles with fine-grained per-label classification (Avg F1 of 0.494), indicating difficulty in distinguishing specific anomaly types without additional context. Adding thermal context features raises binary detection to ANY F1 of 0.885 and ANY AP of 0.964, with a moderate boost to per-label performance (Avg F1 of 0.560). Adding device-history features instead of context features yields the largest single improvement in per-label classification, pushing Avg F1 to 0.671, a relative gain of 19.8% over the episode-plus-context configuration. This result validates the importance of device-relative behavioral comparisons for distinguishing specific anomaly types. The full feature set achieves the best overall performance across both tasks (Avg F1 of 0.730, ANY F1 of 0.889), representing a 47.8% relative improvement in per-label classification over the episode-only baseline.

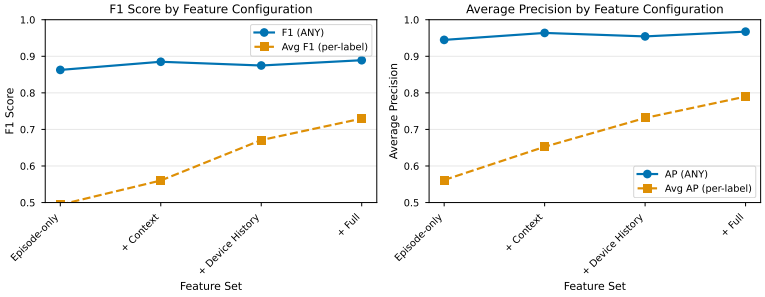



**Fig. 10** Feature ablation study showing progressive performance improvements. Device-history features provide the largest gain for per-label classification (+35.7% Avg F1 relative to baseline). The best overall performance is achieved when combining all three feature groups.

***SHAP feature importance analysis.***

To validate the ablation findings with a model-agnostic measure of feature influence, we compute SHAP (SHapley Additive exPlanations) values [**?** ] for the full XGBoost model. Figure 11 shows the global feature importance rankings based on mean absolute SHAP values across all anomaly types.

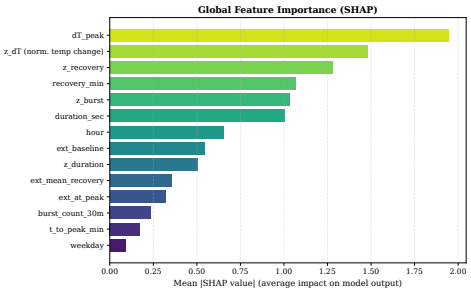



**Fig. 11** Global feature importance based on mean absolute SHAP values. Thermal characteristics and device-relative z-score metrics dominate the model's decisions, consistent with the ablation study findings.

The SHAP analysis confirms and extends the ablation findings. Thermal characteristics, specifically peak temperature deviation (mean |SHAP| = 1.95), recovery time (1.07), and door-open duration (1.01), rank near the top and underpin the episode-only baseline's strong binary detection. Device-relative metrics, including normalized temperature deviation (1.48), normalized recovery time (1.28), and normalized burst intensity (1.03), are also highly influential, validating the ablation finding that historical behavior comparisons are essential for distinguishing fine-grained anomaly subtypes. Context features such as baseline ambient temperature (0.55) and mean recovery ambient temperature (0.36) contribute more moderately but consistently, confirming that environmental context helps to reduce false positives across all anomaly types. The convergence of ablation and SHAP analyses provides strong evidence that the three-layer feature engineering design is well-motivated: thermal characteristics support robust binary detection, device-relative metrics enable fine-grained classification, and context features enhance overall accuracy.

***Per-label performance analysis.***

Figure 12 shows the per-label F1 scores for the full-feature XGBoost model. Thermal anomalies are the easiest to detect. The model achieves F1 of 0.98 for A5 (slow recovery), 0.89 for A4 low-$\Delta T$, and 0.87 for A4 high-$\Delta T$, confirming that anomalous thermal dynamics create distinct, learnable physical signatures in the feature space. Duration and burst anomalies (A2, A3) are detected with moderate but robust performance (F1 of 0.76 and 0.62, respectively). These events are inherently more variable across clinics, making the device-relative z-score features important but inevitably noisier than pure thermal physics. Schedule deviations (A1) represent the most challenging category (F1 of 0.41). This result is expected because off-schedule anomalies are defined entirely by behavioral context and the device-specific operating-hour mask, neither of which is directly encoded as a model feature. Inferring a clinic's unique schedule solely from the timestamps of isolated door openings is the most ambiguous classification task in our taxonomy.

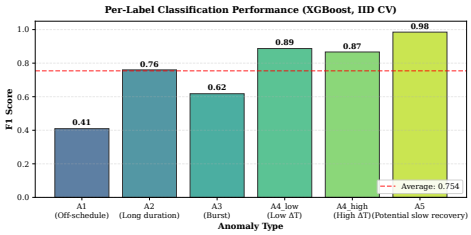



**Fig. 12** Per-label F1 scores across all anomaly types. Thermal anomalies (A5, A4) are classified with the highest accuracy. Behavioral anomalies such as A1 (off-schedule) are the most challenging.

### 4.2.3 Experiment 3: Weak label noise analysis

***Objective.***

Evaluate the robustness of the supervised model to label noise, which is particularly important given that labels are generated by heuristic rules that inherently contain edge cases and logic generalizations. This experiment validates the viability of the weak-supervision approach.

***Methodology.***

We conduct two complementary analyses. In the label perturbation analysis, we systematically corrupt training labels by randomly flipping a fraction $p \in \{0\%, 5\%, 10\%, 15\%, 20\%\}$ of labels and measure performance degradation on a clean test set. In the anomaly co-occurrence analysis, we compute the conditional probability matrix $P(\text{Anomaly}_i \mid \text{Anomaly}_j)$ to understand dependencies between different anomaly types.

***Results and analysis.***

Figure 13 presents the results of both analyses.

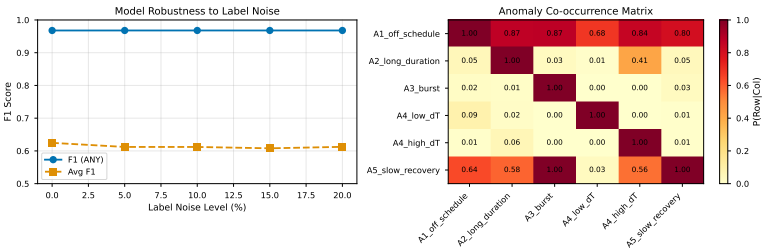



**Fig. 13** Weak label noise analysis. Left: label perturbation shows stable binary detection up to 20% noise, with only minor degradation in per-label precision. Right: anomaly co-occurrence matrix reveals strong dependencies between off-schedule usage (A1), bursts (A3), and slow recovery (A5).

***Label perturbation.***

The model exhibits strong stability under aggressive label noise. Binary detection F1 (ANY F1) remains virtually flat at 0.967 across all noise levels from 0% to 20%, and average precision for binary detection stays constant at 0.994. Per-label F1 shows expected but minor degradation, declining from 0.624 to 0.612 at worst (a maximum relative drop of 2.6% at 15% noise), with no signs of catastrophic failure at any noise level. These results confirm that the model can successfully generalize from weak heuristic labels, tolerating error rates that substantially exceed realistic labeling imperfections. This tolerance is likely attributable to the strong underlying physical signal in the features: even when labels are corrupted, the thermodynamic and behavioral patterns remain informative [**?** ].

***Anomaly co-occurrence.***

The co-occurrence matrix reveals several important dependencies among the anomaly categories. Slow recovery (A5) acts as an absorbing state for thermal and behavioral load, co-occurring with 58% of long-duration events (A2), 100% of burst events (A3), and 56% of high-$\Delta T$ events (A4). Off-schedule access (A1) reliably precedes other anomaly types: 87% of long-duration events, 87% of burst events, and 84% of high-$\Delta T$ events fall within off-schedule periods. Low-$\Delta T$ and high-$\Delta T$ subtypes of A4 show 0% mutual overlap, confirming that the physics-based rules correctly enforce mutual exclusivity. These dependency patterns inform both model design and prediction interpretation: the models leverage continuous features to capture the compound structure of co-occurring anomalies, and the co-occurrence statistics can guide field-level prioritization of alerts.

Together, the label perturbation and co-occurrence analyses provide strong evidence that supervised learning with weak labels is a viable strategy for operational anomaly classification in vaccine cold chains. The model robustly learns the underlying anomaly signatures from heuristic labels without requiring manually curated ground truth or complex noise-filtering procedures.

### 4.2.4 Summary of experimental findings

The three experiments yield four principal findings. First, XGBoost achieves F1 scores between 0.84 and 0.92 and AP scores between 0.93 and 0.98 for binary anomaly detection across all splits, substantially outperforming unsupervised methods and confirming the advantage of supervised learning for this task. Second, the model generalizes well across IID, country-held-out, and temporal splits, with less than a four-percentage-point F1 difference between IID and country splits, indicating robust cross-geography transfer. Third, device-history features (z-scores) contribute a 35.7% relative improvement in per-label F1 compared to the episode-only baseline, and the full three-group feature set yields the best performance for both binary and multi-label classification. Fourth, the model maintains stable binary detection (ANY F1 of 0.967) under up to 20% simulated label noise, with a maximum 2.6% decline in fine-grained classification, demonstrating that weak heuristic supervision is sufficient for training accurate anomaly detectors in this domain.

# 5 Conclusion

This work presented a unified, event-aware framework for operational anomaly analysis in vaccine refrigerators, grounded in millions of real-world telemetry records from 500 devices deployed across 32 countries. By transforming raw sensor streams into semantically structured door-opening episodes annotated with behavioral, thermal, and contextual features, we demonstrated that user interaction patterns and device thermal response can be jointly modeled in an interpretable and scalable manner. The six anomaly categories (A1 through A6) captured distinct modes of operational deviation, and the cross-anomaly analysis revealed that off-schedule access (A1) persistently precedes long-duration openings (A2) and burst episodes (A3), which in turn compound into slow temperature recovery (A5). A multi-label XGBoost ensemble trained on rule-derived weak labels outperformed unsupervised baselines and generalized across countries and temporal splits, while remaining stable under up to 20% simulated label noise. Beyond retrospective detection, the framework provides actionable diagnostics for field engineers by linking each anomaly prediction to interpretable operational factors. Future extensions of this work will target prospective anomaly prediction, where sequence models could anticipate anomalous behavior earlier in the door-opening episode, the integration of richer environmental and supply-chain metadata, and adaptive rule refinement guided by model feedback. Taken together, these results support a shift from reactive, threshold-based alarm systems toward proactive, behavior-aware cold-chain monitoring for global immunization programs.

**Acknowledgements.** This study was conducted using the HPC infrastructure of the University of Luxembourg. We also gratefully acknowledge the support provided by B Medical Systems.

# Declarations

## Funding

This research has been supported by B Medical Systems

## Conflict of interest/Competing interests

Authors declare no conflict of interest.

## Ethics approval and consent to participate

Not applicable

## Consent for publication

All authors read and approved the final manuscript.

## Data availability

The datasets used in this study are not publicly available due to company confidentiality and proprietary restrictions.

## Code availability

The source code developed for this study is not publicly available due to company intellectual property restrictions.

## Materials availability

The materials used in this study are not publicly available due to company confidentiality restrictions.

## Authors' Contributions

A.A.B.: Conceptualization, methodology, programming, data curation, formal analysis, investigation, visualization, and writing—original draft.

N.K., A.K.K.: Methodology, validation, formal analysis, and writing—review and editing.

O.S., S.J., J.A.K., K.L. : Data provision, domain expertise, methodology review-validation, and writing—review and editing.

J.K., T.F.B.: Conceptualization, supervision, project administration, and writing—review and editing.

# References


[1] Kartoglu, U., Ames, H.: Ensuring quality and integrity of vaccines throughout the cold chain: the role of temperature monitoring. Expert Review of Vaccines **21**(6), 799–810 (2022) https://doi.org/10.1080/14760584.2022.2061462 . Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14760584.2022.2061462. Accessed 2025-11-26

[2] Pambudi, N.A., Sarifudin, A., Gandidi, I.M., Romadhon, R.: Vaccine cold chain management and cold storage technology to address the challenges of vaccination programs. Energy Reports **8**, 955–972 (2022) https://doi.org/10.1016/j.egyr.2021.12.039 . Accessed 2025-01-08

[3] Hanson, C.M., George, A.M., Sawadogo, A., Schreiber, B.: Is freezing in the vaccine cold chain an ongoing issue? A literature review. Vaccine **35**(17), 2127–2133 (2017) https://doi.org/10.1016/j.vaccine.2016.09.070 . Accessed 2025-01-07

[4] Centers for Disease Control and Prevention: Vaccine Storage and Handling Toolkit. "Vaccine potency is reduced... Once lost, potency cannot be restored." (2024). https://www.cdc.gov/vaccines/hcp/downloads/storage-handling-toolkit.pdf Accessed 2025-08-21

[5] Das, M.K., Arora, N.K., Mathew, T., Vyas, B., Devi, S.K., Yadav, A.: Temperature integrity and exposure of vaccines to suboptimal temperatures in cold chain devices at different levels in three states of India. Tropical Diseases, Travel Medicine and Vaccines **6**(1), 8 (2020) https://doi.org/10.1186/s40794-020-00109-z . Accessed 2025-01-03

[6] Kitamura, T., Bouakhasith, V., Phounphenghack, K., Pathammavong, C., Xeuatvongsa, A., Kobayashi, A., Norizuki, M., Okabayashi, H., Miyano, S., Mori, Y., Takeda, M., Sugiyama, M., Mizokami, M., Machida, M., Hachiya, M.: Vaccine temperature management in Lao People's Democratic Republic: A nationwide cross-sectional study. Heliyon **7**(6) (2021) https://doi.org/10.1016/j.heliyon.2021.e07342 . Publisher: Elsevier. Accessed 2025-01-03

[7] Chojnacky, M., Rodriguez, A.L.: Effect of thermal ballast loading on temperature stability of domestic refrigerators used for vaccine storage. PLOS ONE **15**(7), 0235777 (2020) https://doi.org/10.1371/journal.pone.0235777 . Accessed 2025-12-03

[8] Habiyaremye, J., Zennaro, M., Mikeka, C., Masabo, E., Jayavel, K., Kumaran, S.: A Data-Driven Predictive Machine Learning Model for Efficiently Storing Temperature-Sensitive Medical Products, Such as Vaccines: Case Study: Pharmacies in Rwanda. Journal of Healthcare Engineering **2021**, 1–12 (2021) https://doi.org/10.1155/2021/9990552 . Accessed 2025-11-28

[9] Martinez-Martinez, E., De La Cruz-Quiroz, R., Fagotti, F., Torres, J.A.: Ambient temperature, refrigerator food load, and door openings effect on a preservation performance indicator based on chicken temperature data and predictive microbiology. Journal of Food Science **88**(9), 3956–3966 (2023) https://doi.org/10.1111/1750-3841.16714 . Accessed 2025-11-28

[10] De Frias, J.A., Luo, Y., Zhou, B., Zhang, B., Ingram, D.T., Vorst, K., Brecht, J.K., Stommel, J.: Effect of door opening frequency and duration of an enclosed refrigerated display case on product temperatures and energy consumption. Food Control **111**, 107044 (2020) https://doi.org/10.1016/j.foodcont.2019.107044 . Accessed 2025-11-28

[11] Belman-Flores, J.M., Pardo-Cely, D., Gómez-Martínez, M.A., Hernández-Pérez, I., Rodríguez-Valderrama, D.A., Heredia-Aricapa, Y.: Thermal and Energy Evaluation of a Domestic Refrigerator under the Influence of the Thermal Load. Energies **12**(3), 400 (2019) https://doi.org/10.3390/en12030400 . Accessed 2025-12-04

[12] Thielmann, A., Puth, M.-T., Kersting, C., Porz, J., Weltermann, B.: Vaccine cold chain in general practices: A prospective study in 75 refrigerators (Keep Cool study). PLOS ONE **14**(11), 0224972 (2019) https://doi.org/10.1371/journal.pone.0224972 . Publisher: Public Library of Science. Accessed 2025-11-26

[13] Lamba, H.K., Sharma, D., Dhir, S., Sushil, S., Ghosh, R.S., Bagchi, S.N., Singh, S., Pooja, P., Kothari, K., Monfardini, E., Doshi, J.: Ensuring Vaccine Temperature Integrity: Monitoring from Storage to Last-Mile Delivery. Global Journal of Flexible Systems Management **25**(3), 559–578 (2024) https://doi.org/10.1007/s40171-024-00401-3 . Accessed 2025-01-03

[14] Biswas, K., Muthukkumarasamy, V., Bai, G., Chowdhury, M.J.M.: A reliable vaccine tracking and monitoring system for health clinics using blockchain. Scientific Reports **13**(1), 570 (2023) https://doi.org/10.1038/s41598-022-26029-w . Publisher: Nature Publishing Group. Accessed 2025-08-11

[15] Hasanat, R.T., Mansoor, N., Mohammed, N., Rahman, M.S., Rasheduzzaman, M.: Development of a monitoring system and power management for an IoT based vaccine carrier. Journal of Physics: Conference Series **1755**(1), 012023 (2021) https://doi.org/10.1088/1742-6596/1755/1/012023 . Publisher: IOP Publishing. Accessed 2025-08-11

[16] Fahrni, M.L., Ismail, I.A.-N., Refi, D.M., Almeman, A., Yaakob, N.C., Saman, K.M., Mansor, N.F., Noordin, N., Babar, Z.-U.-D.: Management of covid-19 vaccines cold chain logistics: a scoping review. Journal of Pharmaceutical Policy and Practice **15**(1), 16 (2022) https://doi.org/10.1186/s40545-022-00411-5 https://www.tandfonline.com/doi/pdf/10.1186/s40545-022-00411-5. PMID: 35236393

[17] Robertson, J., Franzel, L., Maire, D.: Innovations in cold chain equipment for immunization supply chains. Vaccine **35**(17), 2252–2259 (2017) https://doi.org/10.1016/j.vaccine.2016.11.094 . Building Next Generation Immunization Supply Chains

[18] Putrada, A.G., Abdurohman, M.: Anomaly Detection on an IoT-Based Vaccine Storage Refrigerator Temperature Monitoring System. In: 2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), pp. 75–80 (2021). https://doi.org/10.1109/ICICyTA53712.2021.9689150 . https://ieeexplore.ieee.org/document/9689150/?arnumber=9689150 Accessed 2024-08-28

[19] Harrabi, M., Kechiche, L., Hamdi, A., Ouni, B., Hadj Tahar, J.B.: IoT-based Monitoring of Refrigerated Vaccine Storage: A Literature Review and a Proposed Solution. In: 2023 3rd International Conference on Computing and Information Technology (ICCIT), pp. 20–23 (2023). https://doi.org/10.1109/ICCIT58132.2023.10273872 . https://ieeexplore.ieee.org/document/10273872/?arnumber=10273872 Accessed 2024-12-02

[20] Boniol, P., Liu, Q., Huang, M., Palpanas, T., Paparrizos, J.: Dive into Time-Series Anomaly Detection: A Decade Review. arXiv. arXiv:2412.20512 [cs] (2024). https://doi.org/10.48550/arXiv.2412.20512 . http://arxiv.org/abs/2412.20512 Accessed 2025-12-04

[21] Zamanzadeh Darban, Z., Webb, G.I., Pan, S., Aggarwal, C., Salehi, M.: Deep Learning for Time Series Anomaly Detection: A Survey. ACM Computing Surveys **57**(1), 1–42 (2025) https://doi.org/10.1145/3691338 . Accessed 2025-12-04

[22] Chatterjee, A., Ahmed, B.S.: IoT Anomaly Detection Methods and Applications:

A Survey. Internet of Things **19**, 100568 (2022) https://doi.org/10.1016/j.iot.2022.100568 . arXiv:2207.09092 [cs]. Accessed 2025-12-04

[23] DeMedeiros, K., Hendawi, A., Alvarez, M.: A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks. Sensors **23**(3), 1352 (2023) https://doi.org/10.3390/s23031352 . Accessed 2025-12-04

[24] Pai M.M, M., Mayya, V., Bhattacharyya, A., Gottumukkala, S., Saha, D., Das, A.A., Kini, V., Thapliyal, A., Pai, R.M.: A hybrid approach to anomaly detection in industrial multi-sensor data streams. Heliyon **11**(15), 43965 (2025) https://doi.org/10.1016/j.heliyon.2025.e43965 . Accessed 2025-12-04

[25] Audibert, J., Michiardi, P., Guyard, F., Marti, S., Zuluaga, M.A.: USAD: UnSupervised Anomaly Detection on Multivariate Time Series. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3395–3404. ACM, Virtual Event CA USA (2020). https://doi.org/10.1145/3394486.3403392 . https://dl.acm.org/doi/10.1145/3394486.3403392 Accessed 2025-12-04

[26] Jiang, M., Hou, C., Zheng, A., Hu, X., Han, S., Huang, H., He, X., Yu, P.S., Zhao, Y.: Weakly Supervised Anomaly Detection: A Survey. arXiv. arXiv:2302.04549 [cs] (2023). https://doi.org/10.48550/arXiv.2302.04549 . http://arxiv.org/abs/2302.04549 Accessed 2025-12-04

[27] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785 . https://doi.org/10.1145/2939672.2939785

[28] Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17

[29] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: A library for large linear classification. the Journal of machine Learning research **9**, 1871–1874 (2008)

[30] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), 0118432 (2015)

[31] Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)