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

*Research Article*

# Machine Learning Framework for SHM Alarm Log Triage under Alarm Flood and Limited Labeled Data: A Case Study on a Tropical Cable-Stayed Bridge

**Aris Haris Rismayana [1,*], Gatot Sukmara [2], Nana Rachmana Syambas [3], Drees Andriyanto Martono [4], and Acep Purqon [5]**

[1] Department of Informatics Engineering, Politeknik TEDC, Cimahi 40513, West Java, Indonesia; email: rismayana@poltektedc.ac.id
[2] Sub-Directorate of Technical Bridge Construction Planning, Ministry of Public Works, Bandung 40294, West Java, Indonesia; e-mail: gatot.sukmara@pu.go.id
[3] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40132, West Java, Indonesia; e-mail: nanasyambas@itb.ac.id
[4] Research Technology Fiber Optic Solution Society, Institut Teknologi Bandung, Bandung 40132, West Java, Indonesia; e-mail: drees.rtfoss@staff.stei.itb.ac.id
[5] Physics of Earth and Complex Systems, Institut Teknologi Bandung, Bandung 40132, West Java, Indonesia; email: acep.purqon@itb.ac.id
* Corresponding Author : Aris Haris Rismayana

**Abstract:** Structural Health Monitoring (SHM) systems on long-span bridges routinely generate alarm volumes exceeding operators' manual review capacity, a condition known as alarm flood, so that genuine structural anomalies risk being buried among far more numerous instrumentation faults. The problem is acute in developing-country settings, where monitoring falls to small operational teams rather than dedicated staff. This study proposes an alarm-log triage framework that classifies events directly from tabular database records and static sensor metadata, requiring no raw sensor time-series, offering an alternative to the raw-signal paradigm dominating bridge SHM. The dataset comprises 63,934 alarm records logged over 2.5 years on a 641.8-m tropical cable-stayed bridge: labeling collapses as volume rises, leaving only 401 usable labeled records from the pre-flood period, a weakly-supervised task under severe label scarcity. The pipeline integrates alarm-log attributes, causally-ordered temporal window features, and as-built sensor metadata, with fold-internal resampling and recall-oriented operating-point selection. Seven classifiers were benchmarked under identical stratified 5-fold cross-validation, including TabNet alongside linear, kernel-based, tree-ensemble, and neural baselines. Gradient-boosted tree ensembles separate clearly from all other families (F1-Macro 0.7991–0.8860 outside the tree family versus 0.9354–0.9432 within it; McNemar and DeLong tests confirm the separation, $p < 0.05$). Differences among the tree ensembles themselves are not consistently significant, so LightGBM is reported as the selected deployment model on consistent top-tier performance (F1-Macro 0.9432, AUC-ROC 0.9885 at the default threshold) rather than as demonstrably superior. A two-part ablation shows temporal alarm-rate features supply most of the discriminative power, while explicit imbalance handling yields no measurable benefit at the observed ratio, indicating that such mechanisms should be calibrated to imbalance severity rather than applied by default. The results establish operational alarm logs as a viable, lightweight input for structural event triage and document alarm flood quantitatively in a real bridge SHMS.

**Keywords:** Alarm flood; Alarm log classification; Alarm triage; Cable-stayed bridge; Label scarcity; Machine learning; Structural health monitoring; Weakly supervised learning.

**Received: July, 1st 2026**
**Revised: August, 17th 2026**
**Accepted: August, 18th 2026**
**Published: September, 17th 2026**



## 1. Introduction

Cable-stayed bridges are among the most complex assets in national road networks, and monitoring their condition through Structural Health Monitoring Systems (SHMS) has

become an important practice in modern bridge management [1]. These systems integrate multi-type sensor networks that continuously generate high-frequency data [1], [2], and the application of machine learning (ML) to such data has grown rapidly over the past decade [3]–[5].

Despite this growth, full-scale SHMS operation faces a challenge that has received limited research attention: alarm volume can drastically exceed operators' manual review capacity. In the process industries, this condition is known as an alarm flood, in which hundreds to thousands of alarms are triggered within a short period, creating nuisance alarms and impairing the ability to respond appropriately [6], [7]. Although extensively studied in petrochemical and chemical process settings [8]–[10], its occurrence and operational implications in bridge SHM remain inadequately addressed.

The operational consequences in bridge SHM are direct and cumulative. When alarm volume exceeds review capacity, alarms are typically examined in arrival order until operator attention is exhausted, leaving subsequent events unreviewed. Genuine structural anomalies may therefore be buried among more numerous instrumentation faults, while sustained exposure to predominantly spurious alarms can erode operator confidence and contribute to desensitization, as documented in process-industry alarm management [6], [7]. Automated triage addresses this problem by prioritizing alarms so that limited operator attention is directed first toward events most likely to reflect structural rather than instrumental causes, without replacing engineering judgment. This study presents evidence from a real operational case: a cable-stayed bridge whose SHMS recorded tens of thousands of alarms over a 2.5-year monitoring period. As alarm volume increased, the capacity for manual labeling declined substantially, as documented quantitatively in Section 3.5. Database verification identified two clearly defined label classes, System Error (SE) and Abnormal Behavior (AB), providing a basis for binary classification.

This constraint is particularly relevant in developing-country infrastructure contexts. Bridge management in Indonesia has been characterized by persistent challenges in maintenance prioritization and bridge management practices[11], while monitoring duties may be handled by relatively small operational teams. The dataset reflects this setting directly: System Error events substantially outnumber Abnormal Behavior events, so the primary triage burden lies in separating instrumentation-related faults from the relatively rare events warranting engineering attention. Public benchmark alarm-log datasets for bridge SHM are also not readily available because operational alarm records are generally held by bridge authorities and subject to confidentiality and infrastructure-security restrictions.

Prior ML-based anomaly classification in bridge SHM has predominantly used raw sensor time-series, including acceleration, strain, displacement, and vibration data [12]–[15], whereas operational alarm logs as the primary ML input remain largely unexplored. The framework proposed here therefore does not perform anomaly detection on raw signals; instead, it classifies alarm events already recorded in the SHMS database using the alarm record and physical sensor metadata as input features. This study proposes an integrated ML-based alarm triage pipeline rather than a new classifier. Alarm-log attributes, causally ordered temporal features, and static physical sensor metadata are combined into a tabular representation, while class imbalance and decision-threshold selection are evaluated as explicit components of the pipeline. The main contributions of this study are:

1. An operational paradigm using alarm logs as the primary input for bridge SHM event classification, without requiring access to raw sensor time-series.
2. Quantitative documentation of alarm-flood conditions in a real 2.5-year operational bridge SHMS, including their effect on manual labeling capacity.
3. A label-scarcity framework based on operator-generated labels [16] and a curated subset of reliably labeled operational alarm records.
4. An integrated feature representation combining alarm-log, temporal, and physical sensor metadata, together with ablation analysis of the individual feature groups.
5. A comparative evaluation of seven ML classifiers from different algorithmic families, together with component-wise analysis of imbalance handling and decision-threshold selection.

Section 2 reviews related literature. Section 3 describes the bridge SHMS, data sources, alarm generation, dataset construction, and label quality. Section 4 presents the proposed methodology, including feature engineering, model development, imbalance handling, and evaluation.

Section 5 presents and discusses the experimental results, ablation analyses, and comparison with related studies. Section 6 concludes the paper and discusses its limitations and future research directions.

## 2. Literature Review

### 2.1. Machine Learning for Bridge SHM Anomaly Detection

Farrar and Worden [2] established the conceptual framework of SHM as a four-stage process: detection, localization, classification, and damage assessment. Subsequent systematic reviews document the growing dominance of deep learning since 2018 [3], its integration in bridge engineering [4], and the broader use of AI in bridge SHM [5].

For anomaly classification specifically, prior approaches uniformly rely on raw sensor time-series. Samudra et al. [12] proposed a recursive decision tree framework with Random Forest for bridge acceleration data; Kim and Mukhiddinov [13] applied a 1D CNN on cable-stayed bridge data; Abraham et al. [17] proposed unsupervised ensemble fusion of PCA and an autoencoder; Hu et al. [14] developed a hierarchical attention transformer for sensor anomaly detection under limited labels; Zhang et al. [15] applied LSTM to long-span suspension bridge data; and Zhu et al. [18] fused 1D-LBP features with XGBoost on a real cable-stayed bridge.

The most closely related work, by Wang et al. [19], proposes an anomaly detection model for bridge monitoring data that includes an abnormal alarm data evaluation method. Its ML nonetheless operates on raw time-series, with alarm evaluation as a downstream post-processing step. The present study instead uses alarm log records as the primary data source, supplemented only by static as-built sensor metadata. This gap motivates the alarm-log-centric approach proposed here.

### 2.2. Alarm Flood Management in Industry

The ANSI/ISA-18.2 standard [6] defines alarm flood and establishes best practices for alarm management in process industries. Subsequent work has addressed early flood prediction through real-time pattern matching [7], data-driven multivariate nuisance alarm management in chemical processes [8], deep learning for alarm-log-based root cause analysis [20], supervised classification and prediction of industrial alarms [9], and reinforcement learning for proactive flood management [10]. Two distinctions separate bridge SHM from these settings: each event is unique rather than recurring, so a missed structural event carries far greater consequence; and bridge SHM operators are not continuously staffed, making the alarm flood problem more acute.

### 2.3. Imbalanced Data Classification

Class imbalance is a common but often underacknowledged challenge in bridge SHM ML studies [13], [21]. SMOTE [21] is the most widely used oversampling method. Imani et al. [22] confirmed that SMOTE-XGBoost consistently performs best across majority-to-minority ratios from approximately 6:1 to 99:1, spanning moderate to extreme imbalance, and further evidence supports SMOTE-XGBoost for industrial data quality monitoring [23] and optimized XGBoost for imbalanced data [24].

Hybrid resampling extends these approaches: Bonde and Bichanga [25] combined SMOTE with edited nearest neighbours for rare-event detection in financial transactions, illustrating that the choice among SMOTE variants is a design decision rather than a default. Comparative benchmarking across algorithm families is likewise established practice for operational tabular data; Ntayagabiri et al. [26] evaluated a panel of supervised classifiers for IoT attack detection, an application sharing the heterogeneous-feature and event-classification structure of the present task.

Gradient-boosted tree ensembles such as XGBoost [27] and LightGBM [28] are widely reported to perform well on tabular data [29], motivating their inclusion alongside Random Forest [30] as ensemble baselines; Section 4.5 details the rationale for benchmarking these against linear, kernel-based, and neural baselines. Martínez-Heredia and Ventura [16] demonstrated weakly-supervised approaches in industrial predictive maintenance, the framework most directly relevant here, and Deng et al. [31] demonstrated semi-supervised fault diagnosis under limited labels. F1-Macro [32] and AUC-ROC [33] serve as primary evaluation metrics.

## 3. Dataset and System Description

### 3.1. Bridge Overview and SHM System

The research object is a cable-stayed bridge in Indonesia, 641.8 m long with a two-pylon configuration, whose SHMS became fully operational in December 2023. The system comprises 90 sensor channels: 1-axis (7 units), 2-axis (9) and 3-axis (2) accelerometers, KM-100B strain gauges (10 units, ±2,500 $\mu$Strain), cable tension sensors (±5g), GNSS receivers (4 points, 10 Hz), propeller anemometers (1 Hz), and 3D ultrasonic anemometers (20 Hz). Data are acquired via a Gantner Q.station DAQ system at 100 Hz for dynamic sensors.

The operational environment presents compound challenges — high humidity, intense monsoon winds, marine corrosion exposure, and large thermal variation — all of which influence both structural behavior and sensor performance in ways that have received limited attention in the SHM literature [34]–[37]. The system architecture is shown in Fig. 1.

The bridge's identity, precise location, and owning authority are withheld throughout, and all figures and tables omit site-identifying labels — an infrastructure-security measure agreed with the bridge management authority rather than a data-access restriction.

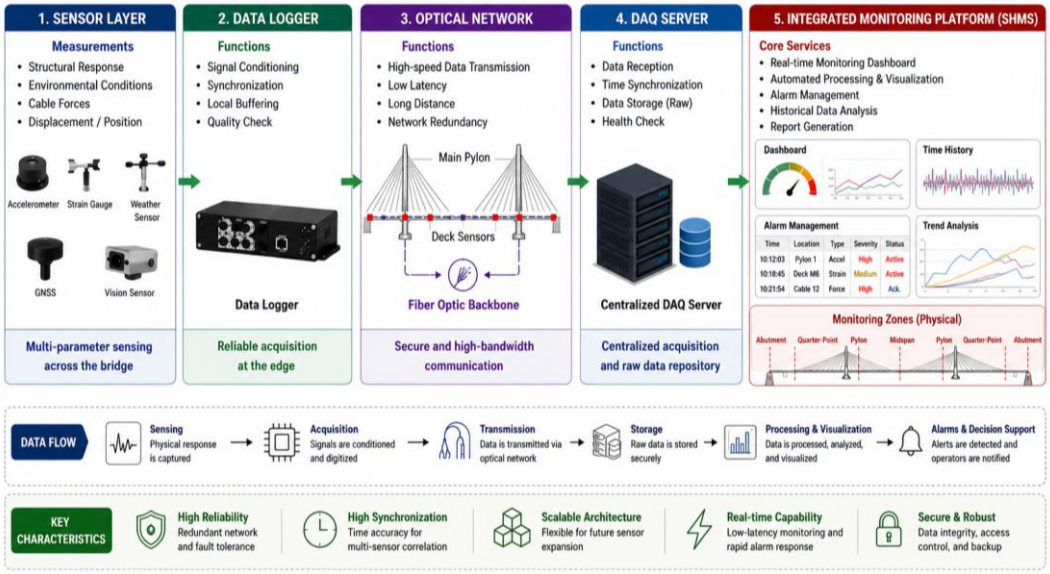



**Figure 1.** SHMS architecture, showing data flow from field sensors through the DAQ server to the integrated monitoring dashboard and the ML-based alarm triage module proposed in this study

### 3.2. Trigger Mode Configuration

The DAQ system implements trigger modes that define when an alarm event is recorded; four active modes are used on this bridge (Table 1). Mode selection reflects the physical characteristics of each sensor type: seismic sensors use absolute thresholds (Type 3) because earthquakes produce sharp spikes; strain gauges and displacement sensors use shift from a 10-minute rolling baseline (Type 6) [38], [39] because deformation manifests as gradual drift; accelerometers require two consecutive threshold-exceeding samples (Type 7) to filter single-sample electronic noise.

Alarm handling operates across two layers. At the machine layer, the DAQ server evaluates each channel against statistically derived thresholds and records the outcome in trigger_lv: Level 1 (warning) for exceedance of a $\mu \pm 2\sigma$ band derived from the normal-operation population, Level 2 (critical) for $\mu \pm 3\sigma$. At the management layer, a three-tier health alert scale — Yellow (operational limit), Orange (potential damage), Red (collapse prevention) — governs field response, from data review within 24 hours to emergency inspection and bridge closure. A manual step connects them: a machine-layer trigger establishes only that a threshold was exceeded, not what caused it, so assigning a health alert level first requires determining whether the event reflects genuine structural behavior or an instrumentation fault. That determination is made by the O&M team through the Abnormal Signal Analysis

Report and is the sole non-automated stage in the chain — exactly the stage the triage task formulated here targets.

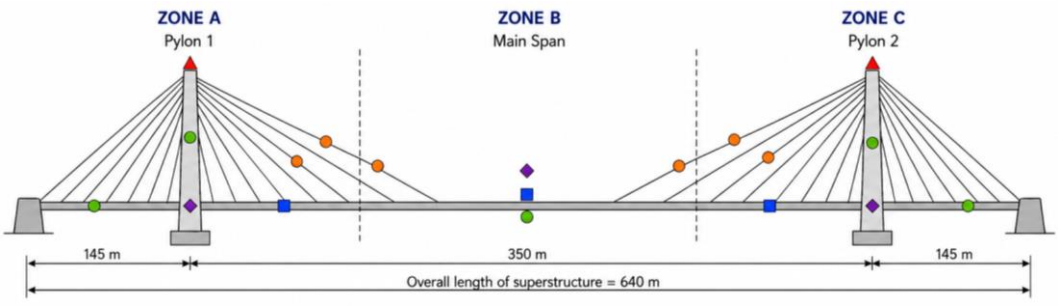
**Table 1.** Trigger Mode Configuration in the Bridge SHM System

| Type | Mode Name | Activation Condition | Channels | Sensor Type |
|---|---|---|---|---|
| Type 0 | OFF / Monitoring | Not activated (passive monitoring) | 22 | Temperature, wind, humidity |
| Type 3 | Max. or Min. Level | Data ≥ Max. OR Data ≤ Min. | 5 | Seismometer (EQK) |
| Type 6 | Shift Level Trigger | (Data − 10-min avg.) ≥ Max. or ≤ Min. | 16 | SSG, joint meter, GNSS N/E |
| Type 7 | Modified Max/Min Level | 2 consecutive data ≥ Max. or ≤ Min. | 47 | Accelerometer |

Monitoring practice for this class of structure is governed nationally. Bridge and tunnel safety provision is regulated under Ministry of Public Works and Housing Regulation No. 10 of 2022 [40], under which the Directorate General of Highways issued Guideline No. 06/P/BM/2025 on the design and installation of bridge SHM systems [41] and Guideline No. 08/P/BM/2025 on their operation and maintenance [42]. The trigger modes in Table 1 are the DAQ vendor's software-layer implementation of the threshold-management requirements those guidelines establish; specific threshold parameters, repair-time targets, and sensor-group priorities are set by the managing authority as site-specific provisions.

### 3.3. Sensor Layout on the Bridge

All 90 sensor positions were confirmed from as-built drawings (Fig. 2) and categorized into five physical zones: Abutment, Cable, Pylon, Quarter-point, and Midspan. Each sensor is additionally classified by its measurement direction as recorded in the SHMS channel configuration, following three coordinate systems used by the DAQ manufacturer: Bridge (BX/BY/BZ — longitudinal/transversal/vertical), used by accelerometers, GNSS, and joint meters; Element (EX/EZ, relative to the mounted element), used by strain gauges and cable tension sensors; and Global (GX/GY/GZ), used by the seismometer. Within the strain-gauge group, direction also encodes mounting location: girder-mounted gauges use EX, abutment-bearing-mounted gauges use BZ. Both metadata items are integrated as ML features [43].




MONITORING ZONES

| | |
|---|---|
| ZONE A Pylon 1 | Pylon, stay cables (west side) and adjacent deck |
| ZONE B Main Span | Main span deck, stay cables (both sides) |
| ZONE C Pylon 2 | Pylon, stay cables (east side) and adjacent deck |

SENSOR INVENTORY SUMMARY

| Sensor Type | Symbol | Quantity | Primary Location |
|---|---|---|---|
| GNSS | ▲ | 4 | Pylons and deck (mid-span) |
| Cable Force Sensor (Tensiometer) | ● | 24 | Stay cables (both sides) |
| Accelerometer | ◆ | 9 | Pylons and deck |
| Strain Gauge | ■ | 10 | Deck (critical sections) |
| Temperature Sensor | ● | 8 | Pylons, deck, and stay cables |

**Figure 2.** Sensor layout from as-built drawings: 90 channels categorized by physical zone and measurement direction, used as Features 13 and 14 in Table 6

### 3.4. Dataset Structure: System Alarm Log

The primary dataset is the log_trigger table in the operational SHMS database, recording every trigger event from December 2023 to June 2026 — 63,934 rows. Database verification confirmed zero duplicate timestamps, indicating serialized processing, with multi-channel simultaneous events grouped into the primary trigger record. Inter-trigger intervals show a mode of approximately 68 seconds, consistent with a systematic DAQ cooldown mechanism. Table 2 describes the main columns used as feature sources.

**Table 2.** Main Columns of log_trigger Used as Feature Sources

| Column | Data Type | Description |
|---|---|---|
| opdatetime | datetime | Trigger event timestamp (second precision, zero duplicate confirmed) |
| channel_id | int | Sensor identity that triggered the alarm |
| value_abnormal | float | Sensor value at trigger time |
| stdv_rate | float | Percentage deviation from configured normal value |
| trigger_lv | int | Alarm level: 1 (warning) or 2 (critical) |
| trigger_type | int | Active trigger mode for that channel (0/3/6/7) |
| class | varchar | Label: System Error or Abnormal Behavior (2 non-NULL values confirmed) |
| main_cause | varchar | Sub-category of cause (NULL in 108 Tier 2 rows) |
| isconfirm | tinyint | Formal counter-verification status (0 = operator-assigned, not yet supervisor-confirmed) |

### 3.5. Alarm Flood Phenomenon and Label Quality Analysis

Labels are assigned by the SHMS O&M team under a codified procedure established by the bridge management authority. Upon threshold exceedance the system issues automatic notifications, archives high-rate raw data for the event window, and requires an Abnormal Signal Analysis Report within 24 hours classifying the cause into exactly the two categories used as class labels here: Abnormal Behavior, for events attributable to external structural loading (heavy vehicles, strong wind, seismic activity, vessel impact); and System Error, for events attributable to the instrumentation and acquisition chain (noise, sensor failure, datalogger failure, power supply faults).

**Table 3.** Trigger volume and labeling rate by operational regime — aggregate counts computed from the operational SHMS database (December 2023 – June 2026; 63,934 alarms over 27 active months)

| Month | Total Triggers | Labeled | Complete | % Label | Notes |
|---|---|---|---|---|---|
| Dec 2023 – Apr 2024 | 391 | 364 | 256 | 93.09% | Normal operation (5 mo.) |
| May 2024 | 2,777 | 145 | 145 | 5.22% | Alarm flood onset |
| Jun – Oct 2024 | 1,643 | 5 | 5 | 0.30% | Chronic alarm flood (4 mo.) |
| Nov 2024 | 18,650 | 5 | 5 | 0.03% | Critical event |
| Dec 2024 – Oct 2025 | 13,681 | 166 | 165 | 1.21% | Chronic alarm flood (9 mo.) |
| Nov 2025 | 4,077 | 257 | 256 | 6.30% | Post-maintenance recovery |
| Dec 2025 – May 2026 | 11,502 | 4 | 3 | 0.03% | Chronic alarm flood (5 mo.) |
| Jun 2026 | 11,213 | 0 | 0 | 0.00% | Critical event |
| Total (27 mo.) | 63,934 | 946 | 835 | 1.48% | 4 mo. system-unavailable |

Note: The full month-by-month breakdown is provided as supplementary material.

Fig. 3 visualizes the inverse relationship between alarm volume and labeling rate. During the normal period volume ranged from 38–120 alarms per month at labeling rates of 82.7–100%. When volume surged in May 2024 (2,777 alarms) the rate dropped to 5.22% and never significantly recovered, reaching 0.03% in November 2024 (18,650 alarms) and 0.00% in June 2026 (11,213 alarms). The single post-flood recovery occurred in November 2025 (6.30%), coinciding with consultant maintenance.

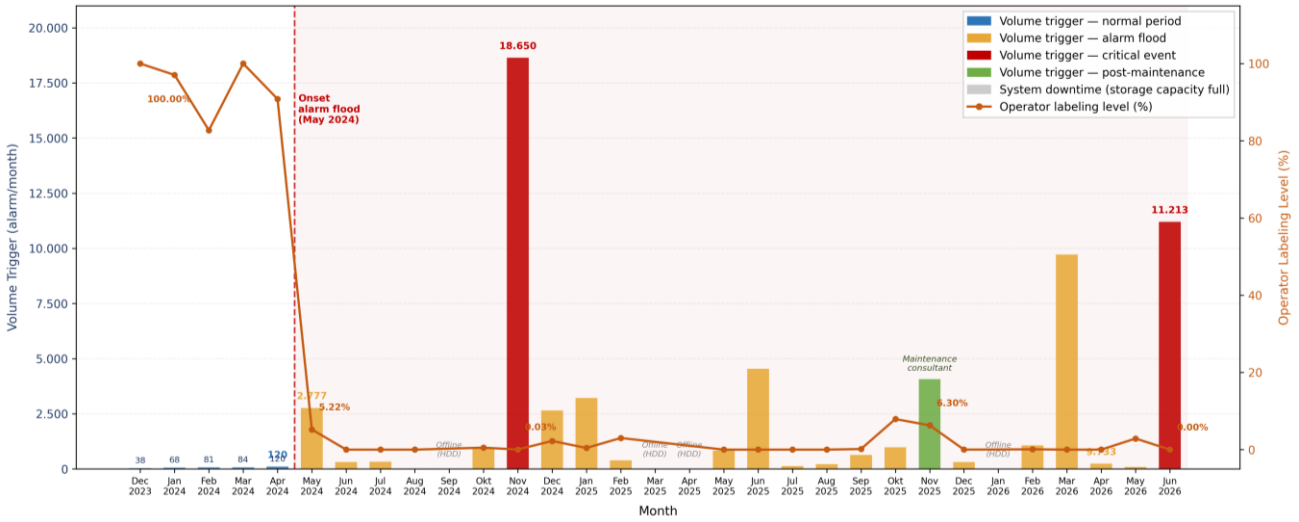



**Figure 3.** Monthly alarm volume (bars, left axis) and operator labeling rate (line, right axis), December 2023 – June 2026, illustrating the collapse of review capacity as volume increases

Of the 63,934 alarms, 946 carry labels (1.48%), of which 835 have both class and main_cause filled. The remaining 111 rows have class but no main_cause: 108 System Error rows, retained as Tier 2 for sensitivity analysis, and 3 Abnormal Behavior rows, excluded from all training sets. The 401 rows from December 2023 – May 2024, hereafter Tier 1, form the high-quality subset used for primary model training (Section 4.1). Database verification confirmed exactly two non-NULL class values — System Error (787 rows) and Abnormal Behavior (159) — and the 62,988 rows with class = NULL are treated as unlabeled data for case illustration (Section 5.5), not as a third class.

All 946 labeled rows have isconfirm = 0, a system field recording whether a supervisor has formally counter-verified the operator's initial assignment. The labels are therefore weak labels [16]: domain-expert assignments made during routine review but not independently confirmed by a second reviewer. This is why a weakly supervised framework is appropriate rather than a fully supervised one.

### 3.6. Label Verification and Main_Cause Distribution

A GROUP BY query on class and main_cause confirmed the hierarchical structure in Table 4. Heavy Wind dominates AB (140 of 156 labeled-cause rows, 89.7%), confirming that most early-period AB events reflected genuine wind-induced structural responses, while Sensor Error dominates SE (513 of 787, 65.2%). The two classes therefore have operationally differentiable and physically interpretable cause profiles.

**Table 4.** Sub-Category Distribution of main_cause Across All Labeled Data — aggregate counts from the operational database.

| Class | Main Cause | Count |
|---|---|---|
| Abnormal Behavior | Heavy Wind | 140 |
| | Others | 8 |
| | Heavy Vehicle | 7 |
| | Site Work | 1 |
| | [main_cause NULL] | 3 |
| System Error | Sensor Error | 513 |
| | [main_cause NULL — Tier 2] | 108 |
| | Noise | 61 |
| | Power Error | 60 |
| | Logger Error | 26 |
| | Others | 19 |

### 3.7. Data Curation and Quality Control

Three dataset-level curation checks preceded any modeling. Database verification found no duplicate timestamps and no malformed or otherwise invalid records in log_trigger (Section 3.4). Records with an unresolved class label were excluded rather than imputed, since the label itself — not a feature — was missing, and imputing it would introduce circular bias. No rows were removed as statistical outliers: extreme values are the primary discriminative signal this study models rather than noise to be suppressed. Sections 3.1–3.7 have characterized the system, dataset, labeling process, and data quality; training-set construction, feature engineering, balancing, and evaluation follow in Section 4 as methodological decisions.

Two derived features used throughout the pipeline (Table 6, Features 5–6) are defined formally here. The z-score normalizes the trigger value against the channel's configured baseline: z_score = (value − stdv_normal) / stdv_normal (1) The baseline shift used for Type 6 (Shift Level) triggers (Section 3.2) is: delta_from_baseline = value − stdv_normal (2) Both quantities are computed from value_abnormal and stdv_rate (Table 2) and are referenced as Eq. (1) and Eq. (2) in the feature descriptions of Table 6.

## 4. Methodology

The methodology is organized as a single pipeline rather than a sequence of independent techniques, and each component addresses a specific characteristic of the dataset described in Section 3: feature engineering responds to the limited discriminative content of an isolated alarm record, imbalance handling to the asymmetric cost of a missed Abnormal Behavior event, and the model comparison protocol to the combination of small sample size, mixed feature types, and moderate imbalance in the Tier 1 set. Fig. 4 summarizes the complete pipeline from sensor acquisition through training-set construction to evaluation.

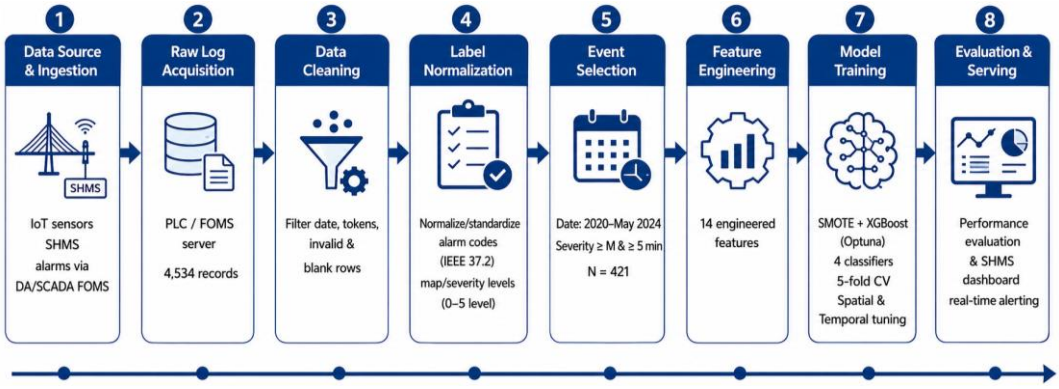



**Figure 4.** End-to-end data pipeline from SHMS sensor acquisition to model evaluation.

Records not meeting the criteria at each stage are excluded from the training pipeline: unlabeled alarms (62,988 rows) are retained for prospective case illustration (Section 5.5); of the 111 rows carrying a class assignment but no main_cause, 108 System Error rows are retained as Tier 2 for the sensitivity analysis (Section 5.4) while 3 Abnormal Behavior rows are excluded from all experiments; months following alarm flood onset are excluded from the primary training set for the reasons given below.

### 4.1. Training Set Selection and Class Distribution (Tier 1 and Tier 2)

The primary training set ("Tier 1") uses 401 rows from December 2023 to May 2024, the six months prior to alarm flood onset. This period is selected because labeling rates were consistently high ($\geq 82.7\%$), all rows have both class and main_cause filled, and it forms a natural boundary before chronic alarm flood began. A secondary pool of 108 rows from subsequent months with class = System Error and main_cause = NULL ("Tier 2") is used only in the sensitivity analysis (Section 5.4); three rows with class = Abnormal Behavior and main_cause = NULL are excluded from all experiments.

Table 5 shows the confirmed class distribution per month in Tier 1. The imbalance ratio is 1.75:1 (SE:AB) — mild [21], [22] and more balanced than the overall labeled set (4.95:1). The high proportion of AB in January and February 2024 (63.1% and 82.6%) likely reflects

major wind events in those months, corroborated by the Heavy Wind dominance in main_cause.

**Table 5.** Class Distribution in the Tier 1 Training Set (n = 401), from a direct database query

| Month | Abnormal Behavior | AB% | System Error | SE% | Total |
|---|---|---|---|---|---|
| Dec 2023 | 9 | 23.7% | 29 | 76.3% | 38 |
| Jan 2024 | 41 | 63.1% | 24 | 36.9% | 65 |
| Feb 2024 | 19 | 82.6% | 4 | 17.4% | 23 |
| Mar 2024 | 10 | 15.4% | 55 | 84.6% | 65 |
| Apr 2024 | 7 | 10.8% | 58 | 89.2% | 65 |
| May 2024 | 60 | 41.4% | 85 | 58.6% | 145 |
| Total | 146 | 36.4% | 255 | 63.6% | 401 |

Two considerations qualify this selection. Tier 1 captures the only period in which the labeling process operated as designed, so it represents the alarm population under unconstrained operator review rather than under collapsed review capacity; its class balance is correspondingly milder (1.75:1 versus 4.95:1 for the full labeled set). This trades temporal coverage for label fidelity, on the reasoning that a model trained on reliably labeled examples of a narrower regime is more defensible than one trained on the sparse, selectively assigned labels of the flood period. Reported performance therefore applies to alarm populations of comparable composition; extension to sustained flood conditions is assessed through the Tier 2 sensitivity analysis (Section 5.4) and remains a limitation in Section 6.

### 4.2. Problem Formulation

The task is binary classification: given one alarm log row, predict System Error (class 0) or Abnormal Behavior (class 1). Database verification (Section 3.5) confirmed exactly two non-NULL class values, and rows with class = NULL are treated as unlabeled rather than as a third class. The setting is weakly supervised in the sense established in Section 3.5 — the labels lack second-reviewer validation, not that they are noisy.

### 4.3. Feature Engineering

Fourteen features were extracted from three sources: the alarm log directly, derived temporal computations, and physical metadata from as-built drawings (Table 6). Window-based temporal features (Features 7–10) were computed using fixed-duration trailing windows applied to the alarm stream sorted by timestamp: a 30-day window for channel- and system-level trigger rate, a 24-hour window for the system-wide burst indicator, and a per-channel time-difference operation for the inter-trigger interval. Each computation at timestamp t uses only records with timestamp ≤ t, ensuring strict causal ordering and eliminating any risk of leakage from future observations.

The hour_of_day feature (Features 12a–b) uses sine–cosine encoding, $\sin(2\pi h/24)$ and $\cos(2\pi h/24)$, preserving cyclical continuity. The delta_from_baseline feature captures the explicit shift magnitude for Type 6 triggers [38], [39]. Physical sensor metadata (Features 13–14) were extracted from as-built drawings and integrated as ML features, following the principle that physical sensor placement carries discriminative information for event classification [43]. One preprocessing step is fold-dependent and therefore specified here rather than in Section 3.7: the two window-based temporal features undefined for a channel's first occurrence were populated with the training-fold median rather than zero, since zero would misrepresent "no prior event" as "immediately repeating"; fewer than 2% of Tier 1 rows required this fill.

### 4.4. Handling Class Imbalance

Despite the mild 1.75:1 ratio, explicit imbalance handling was applied to protect Recall-AB, the operationally critical metric, since a missed structural event is far more costly than an unnecessary inspection. Three complementary strategies were used. First, SMOTE [21] was applied exclusively on training folds within each CV iteration, never on validation folds. Second, class weighting was applied per model where supported: class_weight='balanced' for

Logistic Regression, SVM, and Random Forest; scale_pos_weight=1.75 for XGBoost; is_unbalance=True for LightGBM; the MLP relies on SMOTE alone. Third, the decision threshold was optimized for F1-AB on LightGBM using pooled out-of-fold predictions, raising Recall-AB from 0.9178 to 0.9726 at a modest precision cost (Precision-AB = 0.9342). A methodological limitation applies to this third step: because threshold optimization used the same OOF data as the evaluation, the tuned-threshold metrics are an in-sample optimum rather than an independently validated estimate. A dedicated held-out split was deliberately not carved out of the n = 401 set, since each fold would then retain fewer than 30 Abnormal Behavior cases. Two results with different evidentiary strength are therefore reported: the default-threshold comparison (Table 7), which forms the basis of the model-selection conclusion, and the tuned-threshold operating point (Section 5.1), offered as an operational reference to be validated on an independent holdout in future work. Baseline SMOTE was retained rather than a hybrid variant such as SMOTE-ENN [25], since the 1.75:1 ratio here is mild and the undersampling component of hybrid methods would remove majority-class examples from an already small training pool.

**Table 6.** The 14 named input features (15 feature columns, since hour_of_day is encoded as two cyclic components). Feature names are canonical and used consistently throughout this paper

| # | Feature Name | Source | Type | Justification and Relevance |
|---|---|---|---|---|
| 1 | value_abnormal | log_trigger | Continuous | Sensor value at trigger; absolute anomaly magnitude |
| 2 | stdv_rate | log_trigger | Continuous | Percentage deviation from configured normal baselineimportance #3 (15.5%) |
| 3 | trigger_lv | log_trigger | Ordinal | Statistical extremity marker: Level 1 = μ ± 2σ, Level 2 = μ ± 3σ (set at DAQ layer) |
| 4 | trigger_type | log_trigger | Categorical | Trigger mode (0/3/6/7); reflects sensor type [34] |
| 5 | z_score | Derived | Continuous | (value − stdv_normal)/stdv_normal; cross-sensor normalization |
| 6 | delta_from_baseline | Derived | Continuous | value − stdv_normal; explicit shift magnitude for Type 6 [38], [39] |
| 7 | channel_trigger_rate_30d | SQL | Continuous | Channel trigger frequency over 30 days; identifies chronically malfunctioning sensors (causal) |
| 8 | global_trigger_rate_30d | SQL | Continuous | System-wide trigger rate; alarm flood context indicator — importance #2 (16.6%) |
| 9 | trigger_burst_24h | SQL | Continuous | Alarm volume in 24 h window; detects sudden spikes vs. gradual drift |
| 10 | secs_since_last_same_channel | SQL | Continuous | Inter-trigger interval for same channel importance #4 (6.2%) |
| 11 | cross_channel_flag | Derived | Binary | 1 if another channel alarmed within ±5 min; physical event vs. isolated noise [6], [7] |
| 12a | hour_of_day (sin) | SQL | Cyclic | Cyclic time encoding; sin component (hour_sin) |
| 12b | hour_of_day (cos) | SQL | Cyclic | Cyclic time encoding; cos component (hour_cos) importance #5 (6.0%) |
| 13 | sensor_zone | As-built dwg | Categorical | Physical zone: Abutment / Cable / Pylon / Quarter-point / Midspan [35] |
| 14 | direction_type | As-built dwg | Categorical | Measurement direction: BX / BY / BZ / EX / EZ [39] — importance #1 (33.1%) |

### 4.5. Models and Evaluation Protocol

Seven classifiers were evaluated, each representing a distinct algorithmic family rather than an arbitrary inclusion, compared under an identical preprocessing and validation protocol rather than selected a priori [26]. Two hypotheses are tested. At the family level, axis-aligned tree ensembles are expected to suit this heterogeneous tabular feature space better than distance-based or densely parameterized models, which must impose a common metric on it. At the implementation level, no advantage is assumed between the two gradient-boosting frameworks. Logistic Regression [44] and SVM with an RBF kernel [45] serve as classical

linear and non-linear distance-based baselines, establishing a lower bound. Random Forest [30] is the bagging-ensemble baseline and XGBoost+SMOTE [27] the primary gradient-boosting candidate. LightGBM [28], an independently engineered gradient-boosting framework, is included to assess whether gradient-boosted architectures broadly suit alarm-log tabular data and which implementation is preferable for deployment. A compact two-layer Multilayer Perceptron (32–16 hidden units) represents a lightweight neural baseline, motivated by the documented tendency of deep networks to underperform gradient-boosted trees on small tabular problems [29] — a hypothesis directly testable at n = 401.

Random Forest uses 100 trees with class_weight='balanced'. XGBoost+SMOTE uses scale_pos_weight=1.75 with hyperparameters (100 trees, max depth 6, learning rate 0.1) selected via inner 5-fold CV within each training fold; LightGBM uses the same boosting hyperparameters for direct comparability, with is_unbalance=True. Logistic Regression and SVM use class_weight='balanced' with feature standardization, and the MLP uses two hidden layers (32 and 16 units) with early stopping and standardization. TabNet [46] uses n_d = n_a = 8, n_steps = 3, gamma = 1.3, lambda_sparse = 1e-3, batch_size = 64, and virtual_batch_size = 16 (reduced from defaults of 1024/128 for the ≈320-row training folds), with no eval_set during cross-validation, so early stopping is inactive and no held-out fold can leak into training; training is capped at 100 epochs. Six of the seven models (all except Random Forest) are wrapped in an identical imbalanced-learn pipeline with SMOTE [21] applied exclusively within each training fold, which — combined with the causally ordered temporal features (Section 4.3) — rules out both resampling-induced and temporal leakage. Evaluation uses stratified 5-fold cross-validation maintaining the 1.75:1 distribution in each fold.

The primary evaluation metric is F1-Macro [32], the unweighted average F1 per class, selected because it treats both classes equally regardless of support. Secondary metrics include per-class F1, Precision-AB, Recall-AB, and AUC-ROC [33]. The deployment recommendation rests jointly on F1-Macro and Recall-AB, given the asymmetric cost of false negatives in structural monitoring.

## 5. Results and Discussion

### 5.1. Model Performance Comparison

Table 7 presents the stratified 5-fold CV results for all seven models. Tree-based ensembles occupy the top three positions: LightGBM leads (F1-Macro = 0.9432, AUC-ROC = 0.9885, Recall-AB = 0.9178), ahead of Random Forest and XGBoost+SMOTE, while Logistic Regression and the MLP trail by 6–14 percentage points in F1-Macro. The weak MLP result is consistent with the limited capacity of densely parameterized models to generalize under label scarcity (n = 401) [29].

**Table 7. Model Performance — Stratified 5-Fold CV,** default threshold (Tier 1, n = 401, SE:AB = 1.75:1). Bold = best per metric

| Model | F1-Macro | AUC-ROC | Recall-AB |
|---|---|---|---|
| LightGBM | 0.9432 | 0.9885 | 0.9178 |
| Random Forest | 0.9404 | 0.9855 | 0.9110 |
| XGBoost+SMOTE | 0.9354 | 0.9796 | 0.9178 |
| SVM (RBF) | 0.8860 | 0.9467 | 0.8904 |
| TabNet | 0.8815 | 0.9434 | 0.9041 |
| Logistic Regression | 0.8353 | 0.9202 | 0.8630 |
| MLP | 0.7991 | 0.8968 | 0.9110 |

Note: metrics are computed from pooled out-of-fold (OOF) predictions across the 5-fold procedure, reported as a single aggregate value per model.

These results confirm the family-level hypothesis of Section 4.5, and the mechanism lies in the feature structure documented in Table 6. The input space mixes continuous rate variables with heavy-tailed distributions, an ordinal severity level, and unordered categorical descriptors of sensor placement and orientation. Axis-aligned recursive partitioning accommodates this heterogeneity natively, learning threshold-like rules on each variable in its own scale

without requiring a common metric space. Distance-based methods must impose one, and standardization does not render a five-level zone descriptor commensurable with a 30-day trigger rate. The MLP result reinforces the point from the opposite direction: a densely parameterized approximator has no structural prior favouring axis-aligned splits, and at n = 401 it cannot recover one from data.

TabNet occupies the same non-tree band as the distance- and gradient-based baselines (F1-Macro 0.8815), clearly ahead of the generic MLP but 4.89 percentage points below LightGBM. A native-categorical variant — TabNet's own cat_idxs/cat_dims embeddings with SMOTENC replacing standard SMOTE, averaged over five seeds — scored 0.8823±0.0087, indistinguishable from the protocol-consistent configuration and still roughly seven standard deviations below LightGBM. McNemar and DeLong tests on pooled out-of-fold predictions confirm the separation is not an artifact of a single threshold: LightGBM differs significantly from Logistic Regression, SVM, MLP, and both TabNet configurations (all $p < 0.01$), but not from Random Forest (McNemar $p = 1.00$, DeLong $p = 0.27$), consistent with both sharing the axis-aligned inductive bias. Reconstructed confusion counts show Recall-AB varying little across models (126–134 of 146 AB events recovered) while false positives on System Error range from 9 to 66. The discriminating factor among families is therefore specificity on System Error rather than sensitivity to Abnormal Behavior — operationally relevant, since false positives translate into analyst review load rather than missed events.

Between the two gradient-boosting implementations the evidence does not resolve a preference. McNemar's test on decision-level errors is not significant (p=0.51,9 discordant pairs); DeLong's test on the AUC ranking is significant (p=0.004), but the effect size (ΔAUC=0.0089) is an order of magnitude smaller than the family-level gaps (ΔAUC0.042-0.092), and the component ablation shows XGBoost matching the best LightGBM configuration exactly under class weighting without oversampling (Section 5.2.2). The defensible conclusion is therefore at the family level: gradient-boosted tree ensembles are the appropriate architecture class for alarm-log triage under these conditions, and LightGBM is selected for deployment on consistent top-tier performance across configurations rather than on a demonstrated advantage over XGBoost.

Threshold tuning was subsequently applied to LightGBM, optimizing F1-AB on pooled OOF predictions (Fig. 5a). The optimal threshold (0.383, vs. default 0.5) yields the confusion matrix in Fig. 5b: 245 of 255 System Error and 142 of 146 Abnormal Behavior events are correctly classified, giving Recall-AB = 0.9726, Precision-AB = 0.9342, and F1-Macro = 0.9626 — gains of 1.9 and 5.5 percentage points in F1-Macro and Recall-AB over the default threshold. Given the asymmetric cost of false negatives, LightGBM at threshold = 0.383 is proposed as a provisional operating point. Subject to the in-sample qualification in Section 4.4, it is an operational reference rather than evidence for the model-selection conclusion, which rests on the default-threshold comparison in Table 7.

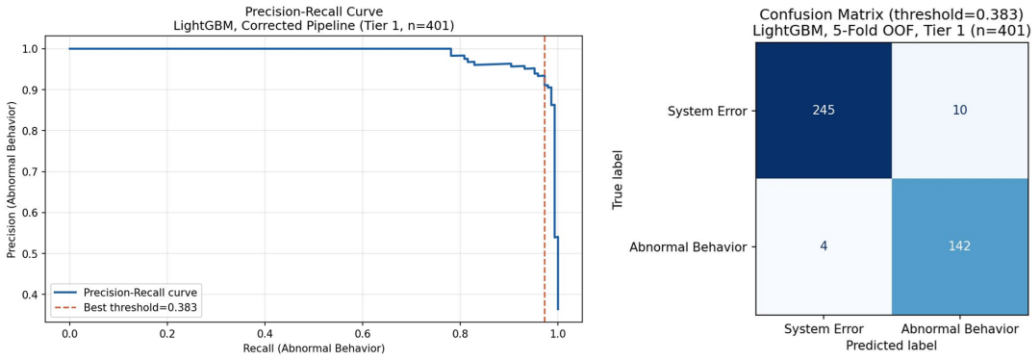



**Figure 5.** LightGBM at the provisional operating point, Tier 1, 5-fold out-of-fold predictions. (a) Precision–Recall curve with the optimal threshold (0.383) marked; (b) confusion matrix at that threshold (n = 401)

In operational terms, the provisional operating point returns 152 of 401 alarms for engineering review, of which 142 are genuine Abnormal Behavior events, 10 are System Error events escalated unnecessarily, and 4 genuine events are missed. This converts a review

burden of 401 items into one of 152 while retaining 97.3% of the events that warrant attention. The unnecessary escalations are inexpensive relative to the alternative: each is resolved by the same instrument check routine maintenance already performs, whereas each missed event forfeits the early-warning function the monitoring system exists to provide.

### 5.2. Ablation Analysis

#### *5.2.1. Feature Groups*

Four feature combinations were evaluated with both XGBoost+SMOTE and LightGBM: (i) alarm-log features alone (6 features); (ii) alarm-log plus temporal features (12 features, adding channel- and system-level 30-day trigger rates, 24-hour burst frequency, time-since-last-trigger, and the cyclic hour-of-day encoding); (iii) alarm-log plus physical metadata (9 features, adding trigger type, sensor zone, and sensor direction); and (iv) the full configuration (15 feature columns, corresponding to the 14 named features in Table 6). trigger_type is grouped with metadata in configuration (iii) despite being recorded in log_trigger, because its value reflects a fixed per-channel DAQ mode configuration (Section 3.2) rather than an event-varying measurement.

Table 8. Ablation Analysis — Contribution of Feature Groups (Stratified 5-Fold CV, Tier 1, n = 401)

| Model | Feature Combination | N Features | F1-Macro | AUC-ROC | Recall-AB |
|---|---|---|---|---|---|
| XGBoost+SMOTE | alarm-log only | 6 | 0.8566 | 0.9348 | 0.8493 |
| LightGBM | alarm-log only | 6 | 0.8819 | 0.9389 | 0.8562 |
| XGBoost+SMOTE | alarm-log + temporal | 12 | 0.9359 | 0.9818 | 0.9384 |
| LightGBM | alarm-log + temporal | 12 | 0.9537 | 0.9874 | 0.9178 |
| XGBoost+SMOTE | alarm-log + metadata | 9 | 0.8707 | 0.9516 | 0.8836 |
| LightGBM | alarm-log + metadata | 9 | 0.8810 | 0.9485 | 0.8904 |
| XGBoost+SMOTE | full feature set | 15 | 0.9354 | 0.9796 | 0.9178 |
| LightGBM | full feature set | 15 | 0.9432 | 0.9885 | 0.9178 |

Table 8 shows that alarm-log features alone yield the weakest performance for both models, confirming that instantaneous trigger characteristics are insufficient to separate System Error from Abnormal Behavior. Adding the temporal group produced the largest gain in both cases (+7 to +8 percentage points F1-Macro), whereas physical metadata alone yielded only a marginal improvement. Notably, the alarm-log+temporal configuration matched — and for LightGBM exceeded (0.9537 vs. 0.9432) — the full configuration, with XGBoost+SMOTE also showing higher Recall-AB under it.

This pattern indicates that the discriminative signal between System Error and Abnormal Behavior originates predominantly from the temporal dynamics of alarm occurrence rather than from static physical attributes of the sensing channel alone. This finding establishes an empirical importance ranking of feature groups: temporal > alarm-log > physical metadata, complementing the single-feature importance analysis in Section 5.3.

#### 5.2.2. Pipeline Components

The remaining components specified in Section 4.4 — synthetic minority oversampling and native class weighting — were evaluated as a factorial combination on the full feature set. Four configurations were run for both gradient-boosting candidates under the same stratified 5-fold protocol at the default threshold: neither mechanism, class weighting alone, SMOTE alone, and both (the pipeline reported in Table 7). Decision-threshold selection is excluded here for the reason given in Section 4.4; its effect is reported separately in Section 5.1 as an operating-point shift.

Neither oversampling nor class weighting improves performance at the imbalance ratio of the Tier 1 set, and both are slightly detrimental in most configurations. For LightGBM the

highest F1-Macro comes from the configuration using neither mechanism (C1, 0.9488), and adding SMOTE lowers it in both weighting states. For XGBoost the best configuration applies class weighting without SMOTE (C2, 0.9488). The effect of class weighting is therefore model-dependent in direction, though marginal in both cases.

**Table 9.** Component-wise ablation of imbalance-handling mechanisms (stratified 5-fold CV, Tier 1, n = 401, full feature set, default threshold). C4 = deployment pipeline; CW = class weighting

| Model | Config | SMOTE | CW | F1-Macro | AUC-ROC | Recall-AB | PR-AUC | F1-Macro (per-fold) |
|---|---|---|---|---|---|---|---|---|
| LightGBM | C1 | off | off | 0.9488 | 0.9880 | 0.9315 | 0.9844 | 0.9487 ± 0.0259 |
| LightGBM | C2 | off | on | 0.9463 | 0.9881 | 0.9384 | 0.9846 | 0.9465 ± 0.0337 |
| LightGBM | C3 | on | off | 0.9432 | 0.9885 | 0.9178 | 0.9832 | 0.9431 ± 0.0220 |
| LightGBM | C4 | on | on | 0.9432 | 0.9885 | 0.9178 | 0.9832 | 0.9431 ± 0.0220 |
| XGBoost | C1 | off | off | 0.9380 | 0.9814 | 0.9178 | 0.9749 | 0.9378 ± 0.0205 |
| XGBoost | C2 | off | on | 0.9488 | 0.9826 | 0.9315 | 0.9775 | 0.9485 ± 0.0182 |
| XGBoost | C3 | on | off | 0.9298 | 0.9795 | 0.9041 | 0.9709 | 0.9297 ± 0.0343 |
| XGBoost | C4 | on | on | 0.9354 | 0.9796 | 0.9178 | 0.9716 | 0.9355 ± 0.0339 |

These differences must be read against the fold-level dispersion in the final column. The spread across all eight configurations (0.9298–0.9488) is 0.019 in F1-Macro, whereas the per-fold standard deviation within a single configuration reaches 0.034. Between-configuration differences are therefore smaller than within-configuration variation, and no configuration is statistically distinguishable from the others. The practical reading is that at 1.75:1 the minority class is already sufficiently represented for gradient-boosted trees to learn a usable decision boundary directly: synthetic samples add little beyond the observed minority instances and, in a pool of 401 rows, appear to introduce mild boundary distortion rather than useful signal.

This finding informs rather than undermines the pipeline reported throughout. The published configuration (C4) applies both mechanisms and sits 0.006 below the best LightGBM variant, well within one fold's standard deviation. The mechanisms are retained deliberately: the Tier 1 ratio reflects a favourable labeling regime, and as labeling coverage extends into flood-period data the effective imbalance is expected to worsen, so retaining them fixes the pipeline across imbalance regimes at negligible cost. Practitioners at comparably mild ratios should nevertheless treat explicit imbalance handling as a choice to verify rather than a default.

One candidate mechanism for the mild detriment under standard SMOTE was checked directly. Because trigger_type, sensor_zone, and direction_type are label-encoded integers, SMOTE's linear interpolation between neighbours can generate synthetic samples with fractional, non-existent category codes, diluting the signal carried by direction_type (Section 5.3). SMOTENC, which draws categorical values from the nearest neighbour instead, was compared against no oversampling and against standard SMOTE for both models. The ordering no_smote ≈ SMOTENC > standard SMOTE held consistently across F1-Macro and Recall-AB, with XGBoost's Recall-AB recovering from 0.9178 to 0.9315. McNemar and DeLong tests were not significant for either model (all $p \geq 0.045$), so categorical interpolation is a directionally consistent partial mechanism rather than a demonstrated cause.

### 5.3. Feature Importance Analysis

Fig. 6 presents LightGBM feature importance (gain-based, normalized) for all 15 feature columns; Table 10 summarizes the top-5 contributors with domain interpretations, using the canonical feature names of Table 6.

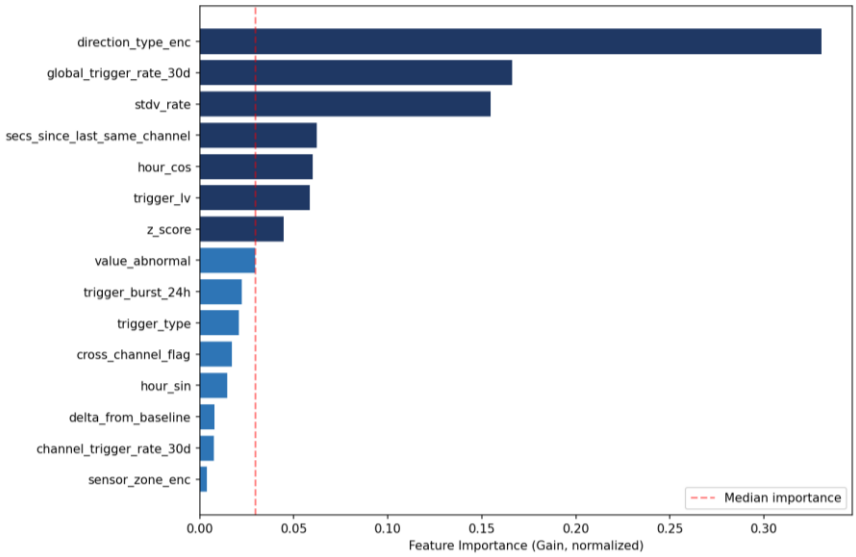



**Figure 6.** LightGBM feature importance (gain, normalized), Tier 1, corrected pipeline. direction_type_enc accounts for 33.1% of total importance; top-5 features together account for 77.3%.

**Table 10.** Top-5 LightGBM Feature Importance (Gain, normalized) and Domain Interpretations.

| Rank | Feature Name (exact as in Table 6) | Importance (Gain) | Domain Interpretation |
|---|---|---|---|
| 1 | direction_type | 0.3305 (33.1%) | Measurement axis encodes a combination of sensor type and mounting location (Section 3.3): EX is used exclusively by girder-mounted strain gauges, which are disproportionately associated with System Error (Event 1, Section 5.5); BX/BY/BZ are shared across accelerometers, GNSS, and abutment-mounted strain gauges, making their association with either class more heterogeneous (Event 2, Section 5.5) |
| 2 | global_trigger_rate_30d | 0.1661 (16.6%) | System-wide alarm-flood context: elevated values indicate systemic SE episodes; low values suggest an isolated event more likely to be genuine AB |
| 3 | stdv_rate | 0.1545 (15.5%) | Percentage deviation from configured normal baseline; large abrupt deviations are more characteristic of sensor faults than gradual structural response |
| 4 | secs_since_last_same_channel | 0.0622 (6.2%) | Repetition interval on the same channel; short intervals indicate persistent, chronic sensor malfunction rather than a one-off structural event |
| 5 | hour_of_day (sin/cos) [cos component] | 0.0600 (6.0%) | Diurnal cyclic pattern; AB events (dominated by Heavy Wind, 89.7%) follow a measurable daily wind cycle, distinct from the non-cyclic pattern of SE events |

The dominance of direction_type as a discriminator (33.1%; Table 10) reflects the channel's actual recorded measurement axis. By contrast, trigger_type, which encodes only the activation mode (Section 3.2), contributes comparatively little—consistent with direction_type subsuming most of the discriminative content that the coarser trigger-mode categorization captures only partially. Notably, sensor_zone ranks lowest among all 15 features (importance < 1%) despite being built from verified as-built drawings: at this sample size the signal is governed by sensor type, orientation, and temporal alarm behavior rather than by location on the bridge. It is retained for physical interpretability rather than predictive contribution.

The dominance of direction_type warrants a generalizability caveat. Its importance reflects the specific failure footprint of sensors on this bridge — girder-mounted strain gauges exhibiting frequent System Error during the study period — rather than a universal link between measurement axis and failure mode. On a bridge where accelerometer channels failed more frequently, the same feature could carry a different or even inverted relationship. This is therefore a site-specific failure signature rather than a transferable rule, reinforcing the single-bridge limitation discussed in Section 6 and the need for retraining on bridges with different instrumentation histories.

### 5.4. Sensitivity Analysis: Tier 1 vs. Tier 1+2

To assess the impact of additional labeled data, LightGBM was also trained on the combined Tier 1+2 set (n = 509, SE:AB = 2.49:1) using the same pipeline, yielding F1-Macro = 0.9491 and AUC-ROC = 0.9896 — both marginally higher than Tier 1 alone — but Recall-AB decreased to 0.9110. All 108 additional rows are System Error, so their inclusion strengthens SE identification and ranking quality at a cost to the minority-class detection rate that is operationally most critical. Tier 1 is therefore retained as the primary training set.

### 5.5. Case Illustration: Critical Events

To illustrate the operational utility of the trained model, LightGBM at threshold 0.383 was applied to the full 63,934-row dataset. These are case illustrations, not formal validation: neither event below has been confirmed by physical field inspection, and the model outputs are predictions from learned feature patterns rather than ground-truth-verified classifications. They demonstrate the interpretability of the model's feature-driven reasoning, not validated real-world performance.

Event 1: SSG 9 burst (November 2024). Sensor SSG 9 (Type 6, Quarter-point zone, EX direction — a girder-mounted strain gauge) generated 15,603 triggers over 17 days with consistently negative (compressive) values. A single-channel prolonged burst with unidirectional values is consistent with sensor mechanical drift or installation interference rather than genuine structural deformation. Applied to all 15,603 triggers, the model classified 100% as System Error (mean P(AB) = 0.0091), matching the reasoning anticipated from trigger_type = Type 6, low secs_since_last_same_channel, and absence of cross_channel_flag. This sensor's EX direction — used exclusively by girder-mounted strain gauges (Section 3.3) — places it in the channel group most strongly associated with System Error (Table 10, Rank 1).

Event 2: SSG 1 Level-2 alarm (June 2026). Sensor SSG 1 at Abutment 1 (an abutment-bearing-mounted strain gauge, BZ direction) recorded 7,482.6 μStrain at 14:39:44 on 13 June 2026 — approximately 50× the configured threshold and exceeding 3× the sensor's physical full scale (±2,500 μStrain), a magnitude far more consistent with sensor failure than with structural deformation. The model classified it as System Error with P(AB) = 0.00096, driven by an extreme z-score (12,663) and a 30-day channel-level trigger rate of 5,336, indicating the channel had been triggering near the system-wide minimum cooldown interval (68 seconds, Section 3.4) for an extended period — persistent malfunction rather than an isolated anomaly. Unlike Event 1, this sensor's BZ direction is shared across accelerometers, GNSS, and abutment-mounted strain gauges, so its classification is better explained by magnitude- and rate-based features than by direction_type.

### 5.6. Comparison with Literature

**Table 11.** Comparison with Representative ML-SHM Studies.

| Study | Input Data | Model | Bridge | Key Distinction |
|---|---|---|---|---|
| Samudra et al. [12] | Raw accel. time-series | Random Forest | Cable-stayed | Supervised, fully labeled, raw signal input |
| Kim & Mukhiddinov [13] | Raw time-series | CNN 1D | Cable-stayed | Data augmentation on raw signal |
| Zhang et al. [15] | Raw time-series | LSTM | Suspension | Anomaly detection on raw sensor signal |
| Wang et al. [19]* | Raw time-series → anomaly detection → alarm evaluation | Transformer-based | Bridge (China) | Alarm classified downstream of raw-signal ML; raw signal still the primary input |
| This study | Alarm log (tabular, no raw signal) | LightGBM (7 baselines) | Cable-stayed (tropical) | Alarm log as primary + only input; alarm triage under flood + label scarcity; 2.5 yr operational data; broader comparative baseline (7 models) + ablation analysis of feature groups and imbalance-handling components |

* Wang et al. [19], Measurement, 2025. The comparison is qualitative: reported metrics are not directly comparable across rows because each study uses a different dataset, bridge, and task definition, so no performance column is included

Table 11 compares this study with representative ML-SHM research. The fundamental distinction is the input data paradigm: prior studies uniformly apply ML to raw sensor time-series, whereas this study applies ML directly to alarm log records — including the closest related work [19], where alarm evaluation sits downstream of raw-signal ML. No raw signal is accessed at any stage here; raw-signal processing occurs only upstream at the DAQ layer (Section 3.2). The approach also carries a practical advantage: feature computation requires only a database query plus a lightweight tabular feature-engineering step on the existing SHMS database, without the storage and processing overhead of raw 100-Hz sensor data. Table 11. Comparison with Representative ML-SHM Studies. * Wang et al. [19], Measurement, 2025. The comparison is qualitative: reported metrics are not directly comparable across rows because each study uses a different dataset, bridge, and task definition, so no performance column is included.

## 6. Conclusions

This study established that operational SHM alarm logs can serve as a practical machine-learning input for structural event triage without requiring access to raw sensor time-series. By treating the alarm record as the unit of analysis, the proposed framework addresses a specific operational bottleneck in bridge SHMS: the rapid growth of alarm volume beyond the capacity of manual review. The results indicate that gradient-boosted tree ensembles are the most suitable model family for the studied alarm-log data, while no statistically conclusive advantage was established among the leading tree implementations. LightGBM is therefore selected as the deployment candidate based on its consistent top-tier performance rather than claimed universal superiority. The feature analysis further shows that temporal alarm dynamics provide the main discriminative information, whereas explicit imbalance handling offers limited benefit under the observed class distribution. These findings suggest that model selection and pipeline design should be driven by the characteristics of operational alarm data rather than by the assumption that more complex or more aggressively balanced models are necessarily preferable.

The study also demonstrates that alarm flooding is not merely a data-volume issue but an operational constraint that directly affects the availability of reliable labels and the ability of personnel to review events. In this context, the proposed framework is best viewed as a triage and prioritization layer within the existing SHMS. It does not replace sensor-based threshold detection or engineering assessment; instead, it helps direct limited operator attention toward alarms that are more likely to warrant further inspection.

Several limitations constrain the interpretation of these findings. The model was trained and evaluated using a small, manually labeled subset of the operational database, and the primary training data represent the pre-flood period in which labeling quality was highest. The tuned decision threshold was derived from the same out-of-fold predictions used for evaluation and therefore requires independent validation before operational adoption. The case-based classifications were also not independently confirmed through field inspection. Finally, the data originate from a single tropical cable-stayed bridge and include a vendor-specific trigger taxonomy, limiting direct generalization to other bridges and acquisition platforms.

Future work should therefore focus on testing transferability and operational readiness rather than simply adding more models. The large pool of unlabeled alarm records can be investigated through semi-supervised learning, while environmental measurements can be incorporated to determine whether they improve discrimination under changing operating conditions. Real-time deployment should be evaluated using stream-based feature computation or incrementally updated state information to replace repeated SQL-based trailing-window queries. Most importantly, validation across multiple bridges and independent operating periods is required to determine whether the alarm-log paradigm remains effective under different sensor configurations, environmental conditions, alarm-generation policies, and levels of alarm flooding.

**Author Contributions:** Conceptualization: A.H.R., G.S., N.R.S., D.A.M., and A.P.; Methodology: A.H.R.; Software: A.H.R.; Validation: A.H.R. and N.R.S.; Formal Analysis: A.H.R., D.A.M., and A.P.; Investigation: A.H.R., D.A.M., and A.P.; Resources: A.H.R., G.S., and D.A.M.; Data Curation: A.H.R. and G.S.; Writing—Original Draft Preparation: A.H.R.; Writing—Review and Editing: A.H.R., G.S., N.R.S., D.A.M., and A.P.; Visualization: A.H.R.;

Supervision: G.S., N.R.S., and D.A.M.; Project Administration: A.H.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data underlying this study are not publicly available, as they originate from the operational SHM database of a national infrastructure asset and are subject to a confidentiality agreement with the bridge management authority. De-identified extracts of the data may be made available by the corresponding author upon reasonable request, subject to the bridge management authority's approval. The methodology and feature engineering pipeline are fully documented in this manuscript for reproducibility.

**Acknowledgments:** The authors thank the bridge management authority for access to the operational SHMS database and as-built drawings. Claude (Anthropic) was used as a writing assistant and argument structuring tool; data analysis, model training, and interpretation were conducted entirely by the research team. One figure was produced with an AI-based diagramming tool from the authors' pipeline specification, with all stage definitions, record counts, and labels verified against the source database.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

[1] L. Sun, Z. Shang, Y. Xia, S. Bhowmick, and S. Nagarajaiah, "Review of Bridge Structural Health Monitoring Aided by Big Data and Artificial Intelligence: From Condition Assessment to Damage Detection," *J. Struct. Eng.*, vol. 146, no. 5, p. 4020073, May 2020, doi: 10.1061/(ASCE)ST.1943-541X.0002535.

[2] C. R. Farrar and K. Worden, "An introduction to structural health monitoring," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 365, no. 1851, pp. 303–315, Feb. 2007, doi: 10.1098/rsta.2006.1928.

[3] J. Jia and Y. Li, "Deep Learning for Structural Health Monitoring: Data, Algorithms, Applications, Challenges, and Trends," *Sensors*, vol. 23, no. 21, p. 8824, Oct. 2023, doi: 10.3390/s23218824.

[4] Z. He, W. Li, H. Salehi, H. Zhang, H. Zhou, and P. Jiao, "Integrated structural health monitoring in bridge engineering," *Autom. Constr.*, vol. 136, p. 104168, Apr. 2022, doi: 10.1016/j.autcon.2022.104168.

[5] R. Zinno, S. S. Haghshenas, G. Guido, and A. Vitale, "Artificial Intelligence and Structural Health Monitoring of Bridges: A Review of the State-of-the-Art," *IEEE Access*, vol. 10, pp. 88058–88078, 2022, doi: 10.1109/ACCESS.2022.3199443.

[6] The International Society of Automation [ISA], *Management of Alarm Systems for the Process Industries*. North Carolina: ANSI (American National Standards Institute), 2016.

[7] M. R. Parvez, W. Hu, and T. Chen, "Real-time pattern matching and ranking for early prediction of industrial alarm floods," *Control Eng. Pract.*, vol. 120, p. 105004, Mar. 2022, doi: 10.1016/j.conengprac.2021.105004.

[8] R. Kaced, A. Kouadri, K. Baiche, and A. Bensmail, "Multivariate nuisance alarm management in chemical processes," *J. Loss Prev. Process Ind.*, vol. 72, p. 104548, Sep. 2021, doi: 10.1016/j.jlp.2021.104548.

[9] S. Sekiou, A. Behloul, R. Nait-Said, and Z. Chiremsel, "Alarms prediction and classification in industrial processes using supervised machine learning techniques: A case study in an Algerian gas plant," *Comput. Chem. Eng.*, vol. 204, p. 109378, Jan. 2026, doi: 10.1016/j.compchemeng.2025.109378.

[10] M. R. Parvez, M. H. Roohi, Z. Guo, and F. Xu, "Proactive management of industrial alarm floods: A reinforcement learning framework for early prediction and operator support," *Control Eng. Pract.*, vol. 161, p. 106341, Aug. 2025, doi: 10.1016/j.conengprac.2025.106341.

[11] S. D. Puspitasari, S. Harahap, and P. Astuti, "A Critical Review of Bridge Management System in Indonesia," in *Proceedings of the 5th International Conference on Rehabilitation and Maintenance in Civil Engineering*, vol. 225, S. A. Kristiawan, B. S. Gan, M. Shahin, and A. Sharma, Eds. Singapore: Springer Nature Singapore, 2023, pp. 381–389. doi: 10.1007/978-981-16-9348-9_33.

[12] S. Samudra, M. Barbosh, and A. Sadhu, "Machine Learning-Assisted Improved Anomaly Detection for Structural Health Monitoring," *Sensors*, vol. 23, no. 7, p. 3365, Mar. 2023, doi: 10.3390/s23073365.

[13] S.-Y. Kim and M. Mukhiddinov, "Data Anomaly Detection for Structural Health Monitoring Based on a Convolutional Neural Network," *Sensors*, vol. 23, no. 20, p. 8525, Oct. 2023, doi: 10.3390/s23208525.

[14] D. Hu, Y. Lin, S. Li, J. Wu, and H. Ma, "Hierarchical Attention Transformer-Based Sensor Anomaly Detection in Structural Health Monitoring," *Sensors*, vol. 25, no. 16, p. 4959, Aug. 2025, doi: 10.3390/s25164959.

[15] J. Zhang, J. Zhang, and Z. Wu, "Long-Short Term Memory Network-Based Monitoring Data Anomaly Detection of a Long-Span Suspension Bridge," *Sensors*, vol. 22, no. 16, p. 6045, Aug. 2022, doi: 10.3390/s22166045.

[16] A. M. Mart\'\inez-Heredia and S. Ventura, "Weak Supervision: A Survey on Predictive Maintenance," *WIREs Data Min. Knowl. Discov.*, vol. 15, no. 2, p. e70022, Jun. 2025, doi: 10.1002/widm.70022.

[17] J. Nesackon Abraham, M. Q. Tran, J. S. Jayaraj, J. C. Matos, M. R. Valluzzi, and S. N. Dang, "Unsupervised Learning-Based Anomaly Detection for Bridge Structural Health Monitoring: Identifying Deviations from Normal Structural Behaviour," *Sensors*, vol. 26, no. 2, p. 561, Jan. 2026, doi: 10.3390/s26020561.

[18] Q. Zhu, W. Li, X. Wang, Q. Zhang, and Y. Du, "Anomaly detection in bridge structural health monitoring via 1D-LBP and statistical feature fusion," *Structures*, vol. 70, p. 107734, Dec. 2024, doi: 10.1016/j.istruc.2024.107734.

[19] L. Wang et al., "Online diagnosis for bridge monitoring data via a machine learning-based anomaly detection method," *Measurement*, vol. 245, p. 116587, Mar. 2025, doi: 10.1016/j.measurement.2024.116587.
[20] N. Javanbakht, A. Neshastegaran, and I. Izadi, "Alarm-Based Root Cause Analysis in Industrial Processes Using Deep Learning," *arXiv*, 2022, doi: 10.48550/arXiv.2203.11321.
[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[22] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.
[23] Y. Han, Z. Wei, and G. Huang, "An imbalance data quality monitoring based on SMOTE-XGBOOST supported by edge computing," *Sci. Rep.*, vol. 14, no. 1, p. 10151, May 2024, doi: 10.1038/s41598-024-60600-x.
[24] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.
[25] L. Bonde and A. K. Bichanga, "Improving Credit Card Fraud Detection with Ensemble Deep Learning-Based Models: A Hybrid Approach Using SMOTE-ENN," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 383–394, Feb. 2025, doi: 10.62411/jcta.12021.
[26] J. P. Ntayagabiri, Y. Bentaleb, J. Ndikumagenge, and H. El Makhtoum, "A Comparative Analysis of Supervised Machine Learning Algorithms for IoT Attack Detection and Classification," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 395–409, Feb. 2025, doi: 10.62411/jcta.11901.
[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[28] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[29] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.
[30] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[31] C. Deng, J. Song, C. Chen, T. Wang, and L. Cheng, "Semi-Supervised Informer for the Compound Fault Diagnosis of Industrial Robots," *Sensors*, vol. 24, no. 12, p. 3732, Jun. 2024, doi: 10.3390/s24123732.
[32] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, p. 268, Dec. 2025, doi: 10.1186/s40537-025-01313-4.
[33] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[34] Y. Ding, X. W. Ye, and Y. Guo, "Data set from wind, temperature, humidity and cable acceleration monitoring of the Jiashao bridge," *J. Civ. Struct. Heal. Monit.*, vol. 13, no. 2--3, pp. 579–589, Mar. 2023, doi: 10.1007/s13349-022-00662-5.
[35] E. Figueiredo, N. Peres, I. Moldovan, and A. Nasr, "Impact of climate change on long-term damage detection for structural health monitoring of bridges," *Struct. Heal. Monit.*, vol. 24, no. 4, pp. 2252–2270, Jul. 2025, doi: 10.1177/14759217231224254.
[36] J. Li, X. Meng, L. Hu, and Y. Bao, "Quantifying the Impact of Environment Loads on Displacements in a Suspension Bridge with a Data-Driven Approach," *Sensors*, vol. 24, no. 6, p. 1877, Mar. 2024, doi: 10.3390/s24061877.
[37] X. He, G. Tan, W. Chu, S. Zhang, and X. Wei, "Reliability Assessment Method for Simply Supported Bridge Based on Structural Health Monitoring of Frequency with Temperature and Humidity Effect Eliminated," *Sustainability*, vol. 14, no. 15, p. 9600, Aug. 2022, doi: 10.3390/su14159600.
[38] J. Kang et al., "Effective alerting for bridge monitoring via a machine learning-based anomaly detection method," *Struct. Heal. Monit.*, vol. 24, no. 6, pp. 3327–3343, Nov. 2025, doi: 10.1177/14759217241265286.
[39] X. Jian, H. Zhong, Y. Xia, and L. Sun, "Faulty data detection and classification for bridge structural health monitoring via statistical and deep-learning approach," *Struct. Control Heal. Monit.*, vol. 28, no. 11, p. e2824, Nov. 2021, doi: 10.1002/stc.2824.
[40] Ministry of Public Works and Housing of the Republic of Indonesia, "Regulation No. 10 of 2022 on the Implementation of Road Bridge and Tunnel Safety." Jakarta, Indonesia, 2022.
[41] Directorate General of Highways, Ministry of Public Works and Housing, Republic of Indonesia, "Guideline for the Design and Installation of Bridge Structural Health Monitoring Systems." Jakarta, Indonesia, 2025.
[42] Directorate General of Highways, Ministry of Public Works and Housing, Republic of Indonesia, "Guideline for the Maintenance of Bridge Structural Health Monitoring Systems." Jakarta, Indonesia, 2025.
[43] Y. Yilmaz, O. C. Bayrak, and A. Soycan, "Evaluation of Machine Learning Algorithms for Classification of Infrastructure Elements in Complex Structures," *KSCE J. Civ. Eng.*, vol. 28, no. 8, pp. 3489–3505, 2024, doi: 10.1007/s12205-024-2062-8.
[44] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.
[45] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[46] S. Ö. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, May 2021, doi: 10.1609/aaai.v35i8.16826.