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.

Reliability-Aware Multivariate Temporal Learning for Early Risk Detection in Cellular Network Infrastructure

Jacob Onu Aji

jacobaji90@gmail.com

MTN Nigeria Communications Plc

Research Article

Keywords: cellular infrastructure, reliability-aware temporal learning, multivariate time series, thermal/electrical failure, early risk detection, predictive maintenance

Posted Date: November 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11471470/v1

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

Additional Declarations: No competing interests reported.

Reliability-Aware Multivariate Temporal Learning for Early Risk Detection in Cellular Network Infrastructure

Jacob Onu Aji

Infrastructure and Mission-Critical Facilities, MTN Nigeria Communications Plc, Nigeria
University of Port Harcourt, Port Harcourt, Nigeria
Corresponding author: jacobaji90@gmail.com | ORCID: 0009-0007-2209-2371

Abstract

Thermal and electrical degradation in cellular network infrastructure develops through coupled changes in equipment temperature, ambient conditions, direct-current (DC) voltage and current, power demand, cooling behaviour and network load. This study develops a reliability-aware multivariate temporal learning framework for forward identification of thermal or electrical abnormal states under sparse and irregular telemetry. A real operational dataset of 12,000 observations from 250 cellular sites and 1,499 assets is analysed with a forward target that excludes contemporaneous alarms and downstream consequences to prevent outcome leakage. The framework separates current-state evidence, recent temporal dynamics and history reliability, and evaluates these components with site-grouped cross-validation. At the 72-h horizon, temporal extreme gradient boosting (XGBoost) achieves a mean area under the receiver operating characteristic curve (ROC-AUC) of 0.5592 ± 0.0295 versus 0.4869 ± 0.0310 for the current-state baseline, with a paired site-bootstrap AUC gain of 0.0741 (95% confidence interval 0.0235–0.1243). A mask-only control reaches 0.5021 and a seven-day rolling-history baseline 0.4745, indicating that sampling availability or longer-window aggregation alone does not reproduce the temporal result. An exploratory learned reliability-gating ablation improves Brier score modestly but reduces ROC-AUC to 0.5220 and is therefore not promoted to the primary learner; instead, history completeness, elapsed gaps and temporal consistency are treated as evidence-quality variables for deployment interpretation. Chronological and leave-one-asset-class-out stress tests further show that generalization remains bounded by the single-environment, sparse-event dataset. The study therefore contributes a reliability-aware evaluation and deployment framework for temporal early-risk identification rather than claiming universal predictive-maintenance capability.

Keywords: cellular infrastructure · reliability-aware temporal learning · multivariate time series · thermal/electrical failure · early risk detection · predictive maintenance

1 Introduction

Cellular infrastructure is a cyber-physical system in which radio equipment, power conversion, backup energy, cooling, environmental control, and network traffic interact. Failure in one subsystem may propagate to other subsystems and result in service degradation, asset damage, or interruption of network availability. For infrastructure operators, the operational value of machine learning therefore depends not only on detecting an abnormal state but on identifying whether recent measurements provide actionable lead time before the abnormal state is observed.

Multivariate time-series anomaly detection has expanded rapidly because the relevant signals are interdependent and their temporal structure can contain weak precursors that are not visible in a single observation. Recent surveys emphasize the importance of temporal dependency modelling, benchmarking discipline, leakage control, and application-specific evaluation for multivariate time-series anomaly detection and predictive maintenance [1–4].

Telecommunication-specific work has also considered machine learning for network anomalies and base-transceiver-station power failures. Ahmed et al. examined multivariate time-series data for BTS power-failure prediction using CNN, LSTM, and hybrid CNN-LSTM models [5]. Cellular-network anomaly detection based on long-term KPI time series has likewise been studied using supervised learning [6]. More recent commercial and patent activity addresses telecommunications equipment failure prediction using environmental measurements and equipment attributes, showing the practical importance of this problem [7]. Recent work in *Telecommunication Systems* also demonstrates the journal's interest in data-driven management of communication systems, including machine-learning-assisted resource monitoring in 5G/edge-RAN environments and learning-based prediction in wireless systems [17, 18].

The unresolved issue addressed here is narrower and more operationally important: whether a heterogeneous, timestamped cellular-infrastructure telemetry history can improve future thermal/electrical event identification while avoiding the common mistake of treating contemporaneous alarms as early-warning variables. The source dataset contains timestamp, site, asset, power, temperature, humidity, HVAC, airflow, radio-load, traffic, and alarm measurements. The previous cross-sectional risk-assessment analysis showed that contemporaneous alarm information dominates observed-state discrimination, while a seven-day future-risk probe remained near chance. This motivates a dedicated temporal study with a forward target and explicit lead-time horizons.

The paper asks six research questions. RQ1: Does temporal history improve future thermal/electrical event discrimination over current-state telemetry alone? RQ2: At which lead times does temporal information provide measurable benefit? RQ3: Does the benefit persist when sequence availability is restricted to observations with complete history? RQ4: How does a recurrent neural baseline compare with a tree-based temporal representation? RQ5: What evidence boundaries are appropriate for deployment claims under sparse site/asset telemetry? RQ6: Is the temporal gain explained by the irregular observation pattern rather than by signal dynamics, and how does it vary across asset classes?

The contributions of the paper are fourfold: (i) a leakage-aware forward-risk formulation for cellular power and thermal infrastructure in which contemporaneous alarms and downstream consequences are excluded by construction; (ii) a reliability-aware temporal representation that explicitly separates instantaneous state, lagged dynamics, elapsed observation intervals and history availability for irregularly sampled telemetry; (iii) a site-level paired-bootstrap evaluation with horizon and history-completeness sensitivity, supplemented by chronological and leave-one-asset-class-out stress tests; and (iv) controlled robustness analyses using a mask-only sampling-pattern model, a seven-day rolling-history baseline, an exploratory reliability-gating ablation, and asset-class event stratification. The resulting framework treats temporal history not as universally trustworthy input, but as evidence whose usefulness depends on sampling completeness and operational context.

2 Related Work

Multivariate time-series anomaly detection methods generally exploit temporal dependencies, cross-variable interactions, or both. Correia et al. [1] survey online model-based approaches and emphasize the need for realistic evaluation and representative datasets. Wang et al. [2] review deep anomaly detection specifically for multivariate time series, while Jia et al. [3] organize deep time-series anomaly detection families and evaluation issues. Syed et al. [4] review time-series algorithms in predictive maintenance and identify anomaly detection, failure classification, and prognostics as related but distinct tasks.

In telecommunications, Ahmed et al. [5] provide a particularly relevant precedent by studying multivariate time-series prediction of BTS power failure. Their work demonstrates the applicability of CNN, LSTM [10], and CNN-LSTM models to power and environmental monitoring. The present study differs in target construction and evaluation: it combines thermal and electrical abnormal states, uses a site-grouped validation unit, excludes contemporaneous alarms from the future-risk experiment, and evaluates a sequence-derived temporal representation against current-state models using paired site-level uncertainty. The journal also has a sustained reliability and systems-management perspective, including work on reliable communication-system design [16], machine-learning-assisted resource monitoring for 5G systems [17], and LSTM-based data-driven estimation in telecommunications [19].

The distinction between contemporaneous classification and genuine early warning is central. An operational alarm may be generated because the system has already crossed a threshold. Including it in a model whose target is the same or downstream event can produce high apparent discrimination without meaningful lead time. The present paper therefore treats alarm exclusion as a design requirement for the forward-risk task. This follows the broader methodological warning in multivariate anomaly-detection literature that evaluation protocols can materially change apparent performance [1–4].

3 Dataset and Problem Formulation

The source dataset contains 12,000 real telemetry observations from 250 cellular sites and 1,499 assets spanning six infrastructure classes. The observation period is approximately ten weeks, from January to March 2026. The asset classes are HVAC, Generator, UPS/Rectifier, DC Rectifier, RF Power Amp, and Transfer Switch. The analyzed fields contain no missing values in the source workbook. The dataset includes 471 overheating events, 47 voltage anomalies, and 9 high-humidity events; the present paper excludes high-humidity from the primary outcome in order to focus specifically on thermal/electrical failure states.

Table 1 summarises the dataset and the outcome construction, and Fig. 1 shows the event composition before and after forward labelling.

Table 1 Dataset characteristics and outcome construction

Characteristic	Value
Raw observations	12,000
Cellular sites / assets	250 / 1,499
Asset classes (6)	HVAC, Generator, UPS/Rectifier, DC Rectifier, RF Power Amp, Transfer Switch
Observation period	≈ 10 weeks (January–March 2026)
Missing values in analysed fields	None

Characteristic	Value
Median site-level observation gap	≈ 23.75 h (irregular sampling)
Recorded events (all rows)	Overheating 471; Voltage anomaly 47; High humidity 9 (excluded from outcome)
72-h eligible set $\lvert\mathcal{D}_{72}\rvert$	9,991 observations
72-h future events	156 (Overheating 147; Voltage anomaly 9); prevalence π_{72} = 1.56%
Validation unit	Site_ID (5-fold GroupKFold)

The thermal/electrical event label is defined from the recorded event field as:

$$Y_i = \begin{cases} 1, & \text{if } E_i \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases}, \quad \mathcal{E} \triangleq \{\text{Overheating}, \text{Voltage anomaly}\} \tag{1}$$

The primary early-risk task is forward prediction. For a reference observation at time t and a horizon h, the future label is one if at least one thermal/electrical abnormal state is observed for the same asset in the interval (t, t+h]. Rows without a complete follow-up interval are excluded. The current observation itself is never counted as a future event.

$$Y_{i,t+h} = \mathbb{1}\left[\exists \tau \in (t, t+h] : E_{i,\tau} \in \mathcal{E}\right] \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Let T_{end} denote the last timestamp of the observation window. The eligible set at horizon h and the corresponding forward-event prevalence are

$$\mathcal{D}_h = \{(i,t) : t + h \leq T_{\text{end}}\} \tag{3}$$

$$\pi_h = \frac{1}{|\mathcal{D}_h|} \sum_{(i,t)\in\mathcal{D}_h} Y_{i,t+h}, \quad \pi_{72} = \frac{156}{9{,}991} \approx 0.0156 \tag{4}$$

Eq. (3) enforces a complete follow-up interval, and Eq. (4) defines the no-skill reference against which PR-AUC is interpreted. The labelling scheme is illustrated in Fig. 2.

At the 72-h horizon, 9,991 observations are eligible and 156 have a subsequent thermal/electrical event, yielding a 1.56% future-event prevalence. Overheating accounts for almost all of these future events (147), whereas future voltage anomalies are much rarer (9); therefore, event-specific voltage modelling is treated as exploratory and not used for a standalone high-confidence claim.

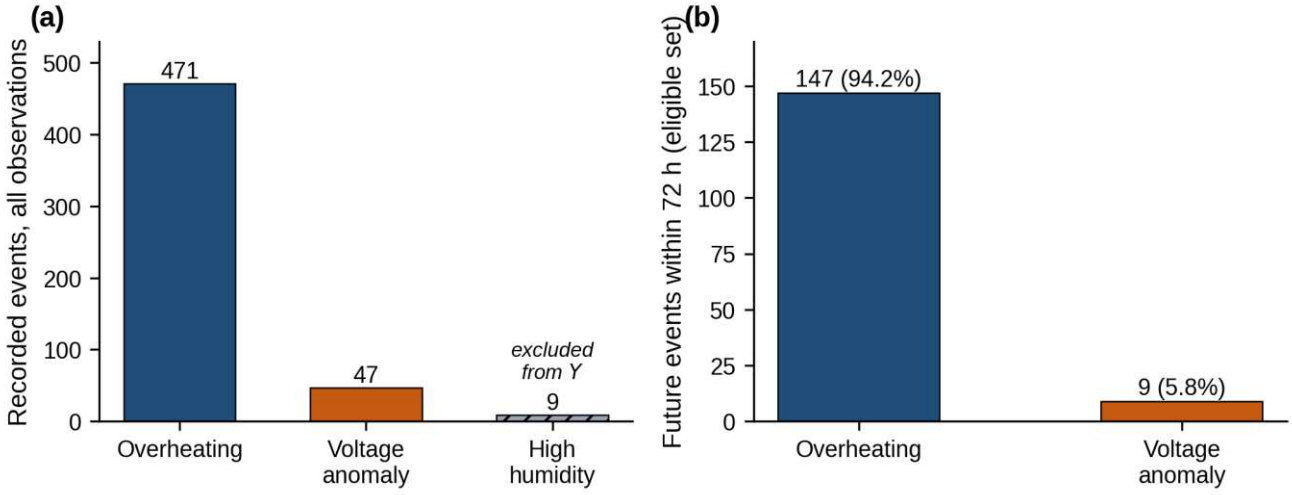


Fig. 1 Event composition. (a) Recorded events across all 12,000 observations; high-humidity events are excluded from the outcome. (b) Future thermal/electrical events in the 72-h eligible set, showing the dominance of overheating

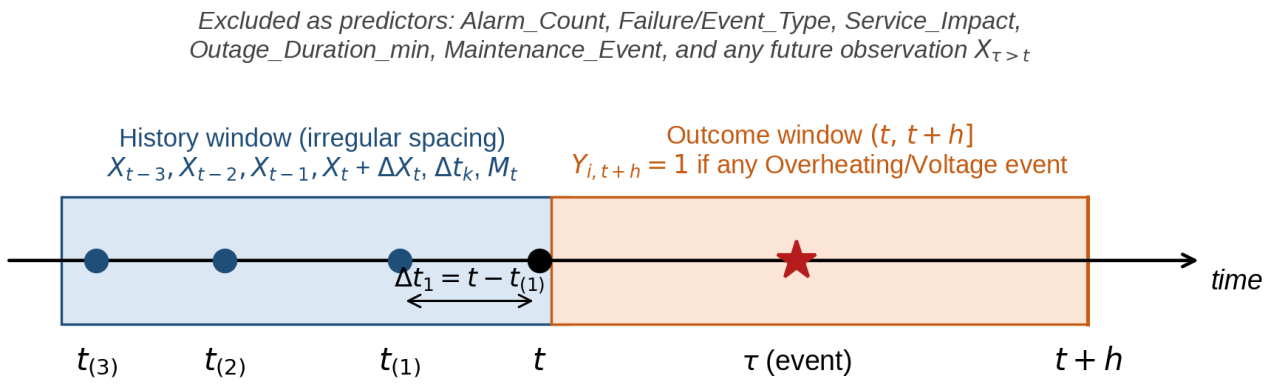


Fig. 2 Forward-labelling and history construction for asset i at reference time t. The history window contains the current observation and up to three irregularly spaced previous observations; the outcome window (t, t+h] determines Y_{i,t+h} (Eq. 2). Contemporaneous alarms and downstream consequences are never used as predictors

4 Multivariate Temporal Representation

For each asset, observations are ordered by timestamp. The temporal representation uses the current engineered telemetry vector together with up to three previous observations. This design preserves a fixed dimensional representation while allowing irregular observation intervals to be represented explicitly. The current state includes electrical, thermal, environmental, HVAC, airflow, RF-load, PRB-utilization, traffic, and user variables, together with physics-derived quantities.

$$P_{\mathrm{VI}} = V_{\mathrm{DC}} \cdot I_{\mathrm{DC}} \tag{5}$$

$$r_{\mathrm{P}} = P_{\mathrm{DC}} - P_{\mathrm{VI}} \tag{6}$$

$$\Delta T_{\mathrm{EA}} = T_{\mathrm{equipment}} - T_{\mathrm{ambient}} \tag{7}$$

$$\Delta T_{\mathrm{HVAC}} = T_{\mathrm{return}} - T_{\mathrm{supply}} \tag{8}$$

$$L_{\mathrm{th}} = \frac{P_{\mathrm{DC}}}{\max(\mathrm{Airflow}, \varepsilon)}, \quad \varepsilon > 0 \tag{9}$$

Here P_{VI} is the electrical power implied by the measured DC voltage and current, r_{P} is the residual between reported and implied DC power, ΔT_{EA} is the equipment-to-ambient temperature rise, ΔT_{HVAC} is the HVAC return–supply differential, and L_{th} is a thermal-loading proxy expressing DC power per unit airflow; ε is a small positive constant preventing division by zero.

For lag k, the model receives X_{t−k}, where the subscript t−k denotes the kth previous observation of the same asset (recorded at time t_(k)) rather than a fixed-step lag, because sampling is irregular. With d engineered current-state variables and d_Δ differenced variables, the temporal vector is

$$Z_t = [X_t, \tilde{X}_{t-1}, \tilde{X}_{t-2}, \tilde{X}_{t-3}, \Delta X_t, \Delta t_1, \Delta t_2, \Delta t_3, M_t] \in \mathbb{R}^{4d+d_\Delta+6} \tag{10}$$

where ΔX_t is the current-to-previous change for the engineered numeric variables, Δt_k is the elapsed time since the kth previous observation, and M_t records the availability of each historical step. Missing lag values are imputed from the training-fold feature distribution and the lag-availability indicators are retained. This is necessary because the source telemetry is irregular rather than a perfectly sampled panel.

Formally, let $n_i(t)$ be the number of previous observations of asset i before t, and let $t_{(k)}$ be the timestamp of the kth previous observation. The change, gap, mask and imputation operators are

$$\Delta X_t = X_t^{(\Delta)} - X_{t-1}^{(\Delta)} \tag{11}$$

$$\Delta t_k = t - t_{(k)}, \quad k \in \{1, 2, 3\} \tag{12}$$

$$M_{t,k} = \mathbb{1}[n_i(t) \geq k], \quad M_t = (M_{t,1}, M_{t,2}, M_{t,3})^{\top} \tag{13}$$

$$\tilde{X}_{t-k,j} = M_{t,k} X_{t-k,j} + (1 - M_{t,k}) \hat{\mu}_j^{(\mathrm{tr})} \tag{14}$$

where the superscript (Δ) restricts to the engineered numeric variables and $\hat{\mu}_j^{(\mathrm{tr})}$ is the imputation statistic of feature j estimated on the training fold only. Because $M_{t,k}$ is retained in Z_t, the learner can distinguish an observed lag value from

an imputed one; the consequence of this design is examined in Section 7.3. The principal notation is summarised in Table 2.

Table 2 Principal notation

Symbol	Definition
i, s(i), t	Asset index, its site, and reference timestamp
h	Prediction horizon (24, 48, 72, 96, 168 h)
E_{i,t}, $\mathscr{E}$	Recorded event type; target set {Overheating, Voltage anomaly}
Y_i, Y_{i,t+h}	Contemporaneous label (Eq. 1); forward label (Eq. 2)
$\mathcal{D}$_h, π_h	Eligible set with complete follow-up (Eq. 3); forward-event prevalence (Eq. 4)
X_t, X_{t−k}	Current engineered telemetry vector (dimension d); kth previous observation
ΔX_t, Δt_k, M_{t,k}	Current-to-previous change; elapsed time to kth previous observation; lag-availability indicator
n_i(t), t_(k)	Number of previous observations of asset i before t; timestamp of kth previous observation
Z_t	Temporal representation (Eq. 10)
p̂, σ(·)	Predicted event probability; logistic function
f_m, M, Ω	Regression tree m, ensemble size (250), tree-complexity penalty
$\mathcal{P}$, $\mathcal{N}$, n_+, n_-	Held-out positive and negative sets and their sizes
$\mathcal{S}$, B	Set of sites ($\lvert\mathcal{S}\rvert$ = 250); number of bootstrap replicates
Δ_AUC, Λ_h	Paired ROC-AUC difference (Eq. 31); PR-AUC lift over prevalence (Eq. 35)

4.1 Reliability-aware temporal evidence framework

The temporal representation is interpreted through three distinct evidence channels: current-state evidence, temporal-dynamics evidence and history reliability. Let $p_C(t)$ and $p_T(t)$ denote the current-state and temporal risk estimates for the same reference observation. The temporal increment on the log-odds scale is

$$\Delta_T(t) = \operatorname{logit} p_T(t) - \operatorname{logit} p_C(t), \quad \operatorname{logit}(p) = \ln \frac{p}{1-p} \tag{15}$$

Rather than assuming that $\Delta_T(t)$ is equally trustworthy for every observation, the framework defines the evidence-quality vector

$$H_t = [n_t, \Delta t_1, \Delta t_2, \Delta t_3, C_t], \quad n_t = n_i(t) \in \{0,1,2,3\} \tag{16}$$

where n_t is the number of available previous observations and C_t is a temporal-consistency statistic. With $q_j, j = 1, \ldots, J$, the physics-informed quantities of Eqs. (5)–(9) and $\dot{r}_{j,k}$ their rate of change between consecutive observations (with $t_{(0)} = t$ and $q_{j,(0)}$ the current value),

$$\dot{r}_{j,k} = \frac{q_{j,(k-1)} - q_{j,(k)}}{t_{(k-1)} - t_{(k)}}, \quad C_t = \frac{1}{J}\sum_{j=1}^{J} \left| \frac{1}{n_t} \sum_{k=1}^{n_t} \operatorname{sgn} \dot{r}_{j,k} \right| \in [0,1] \tag{17}$$

so that $C_t = 1$ when every physics-informed quantity moves consistently in one direction across the available history and $C_t \to 0$ when successive changes alternate in sign; C_t is set to zero when $n_t < 2$. A reliability score and the corresponding deployment rule are

$$R_t = \rho(H_t) \in [0,1], \quad \hat{p}_t = \begin{cases} p_T(t), & R_t \geq r^* \\ p_C(t) + \text{engineering review}, & R_t < r^* \end{cases} \tag{18}$$

where ρ is any monotone map of the evidence-quality vector and r^* is an operator-chosen threshold: high R_t supports use of temporal evidence, whereas low R_t triggers current-state interpretation and engineering review. This distinction matters because the history-sensitivity experiment shows that the aggregate temporal gain attenuates as more complete histories are required. Reliability is therefore treated as an evidence-quality layer rather than an assumption that a more complex temporal learner is automatically more general.

An exploratory learned gate was also tested as a direct implementation of this idea. The gate interpolates current-state and temporal log-odds using history completeness, elapsed gaps, temporal consistency and physics-informed stress indicators S_t:

$$\operatorname{logit} p_G(t) = g_t \operatorname{logit} p_T(t) + (1 - g_t) \operatorname{logit} p_C(t), \quad g_t = \sigma(w^\top [H_t, S_t] + b) \tag{19}$$

where $g_t \in (0,1)$ is learned on training folds only. Because the gate reduced pooled ROC-AUC relative to the locked temporal learner (Section 7.8), it is reported as a negative ablation and is not used to replace the primary model. This design

choice prevents architectural novelty from being claimed solely on the basis of a more elaborate model when the available event population does not support a performance improvement.

5 Leakage-Aware Early-Risk Learning

The forward-risk task excludes contemporaneous Alarm_Count because it is a direct operational indicator and can occur at or immediately before the target event. Failure/Event_Type, Service_Impact, Outage_Duration_min, Maintenance_Event, and future observations are also excluded. Site_ID is used only for grouping and never supplied as a predictive feature. Calibration [13] and all learned preprocessing operations, implemented with scikit-learn [14], are restricted to training folds.

Table 3 lists every variable group together with its role in the forward-risk task.

Table 3 Variable groups and leakage status in the forward-risk task

Group	Variables	Role
Electrical	DC voltage, DC current, DC power	Predictor (current + lags)
Thermal	Equipment temperature, ambient temperature	Predictor (current + lags)
Environmental / cooling	Humidity, HVAC supply and return temperature, airflow	Predictor (current + lags)
Network load	RF load, PRB utilisation, traffic, users	Predictor (current + lags)
Physics-derived	P_VI, r_P, ΔT_EA, ΔT_HVAC, L_th (Eqs. 5–9)	Predictor (current + lags)
Temporal	X_{t−k}, ΔX_t, Δt_k, M_{t,k} (Eqs. 10–14)	Temporal models only
Alarm_Count	Contemporaneous alarm count	Excluded: direct operational indicator
Failure/Event_Type	Recorded event field	Excluded: defines the target
Service_Impact, Outage_Duration_min	Downstream consequences	Excluded: outcome leakage
Maintenance_Event	Intervention record	Excluded: outcome leakage
Future observations	Any X_τ with τ > t	Excluded: look-ahead
Site_ID	Site identifier	Grouping variable only

Three model families are compared. The current-state baseline is XGBoost trained only on the current observation and serves as the non-temporal telecommunication-system control. The principal temporal learner is XGBoost trained on the multivariate lag-and-delta representation. A small gated recurrent unit (GRU) network is used as a recurrent benchmark with the same four-step history and explicit mask/time-gap channels.

The temporal XGBoost learner [8] estimates the forward-event probability with an additive ensemble of regression trees trained by regularised logistic loss:

$$\hat{p}_{i,t+h} = \sigma\left(\sum_{m=1}^{M} f_m(Z_t)\right), \quad f_m \in \mathcal{F}, \quad M = 250 \tag{20}$$

$$\mathcal{L} = \sum_{n\in\mathcal{I}_{\mathrm{tr}}} \ell(Y_n, \hat{p}_n) + \sum_{m=1}^{M} \Omega(f_m), \quad \ell(y,p) = -[y \ln p + (1-y)\ln(1-p)] \tag{21}$$

$$\Omega(f) = \gamma J_f + \frac{1}{2}\lambda \sum_{j=1}^{J_f} w_j{}^2, \quad \lambda = 1 \tag{22}$$

where $\mathcal{F}$ is the space of regression trees, $\mathcal{I}_{\mathrm{tr}}$ indexes training-fold observations, J_f is the number of leaves of tree f, w_j are its leaf weights, and $\lambda = 1$ is the L2 penalty. The current-state baseline uses the same model with Z_t replaced by X_t.

The GRU benchmark [9] processes the input sequence $u_k = [\tilde{X}_{t-(4-k)}, M_t, \Delta t_\bullet]$, $k = 1, \ldots, 4$, ordered from oldest to current, through

$$\begin{aligned} z_k &= \sigma(W_z u_k + U_z h_{k-1} + b_z) \\ r_k &= \sigma(W_r u_k + U_r h_{k-1} + b_r) \end{aligned} \tag{23}$$

$$\begin{aligned} \tilde{h}_k &= \tanh(W_h u_k + U_h(r_k \odot h_{k-1}) + b_h) \\ h_k &= (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k \end{aligned} \tag{24}$$

$$\hat{p}^{\mathrm{GRU}} = \sigma\left(w_o^{\top}\varphi(W_d h_4 + b_d) + b_o\right), \quad h_k \in \mathbb{R}^{32}, \quad W_d \in \mathbb{R}^{16\times 32} \tag{25}$$

where $\odot$ denotes the element-wise product, φ is the activation of the 16-unit dense layer, and all weights are learned on training folds by minimising the binary cross-entropy ℓ in Eq. (21).

The XGBoost configuration uses 250 estimators, maximum depth 3, learning rate 0.05, subsample 0.9, column subsampling 0.9, min_child_weight 2, L2 regularization 1, histogram tree construction, and random seed 42. The GRU uses 32 hidden units, a 16-unit dense layer, Adam optimization with learning rate 0.001, weight decay 10^-4, batch size 256, and 25 training epochs. The recurrent model is included as a benchmark rather than as the central contribution.

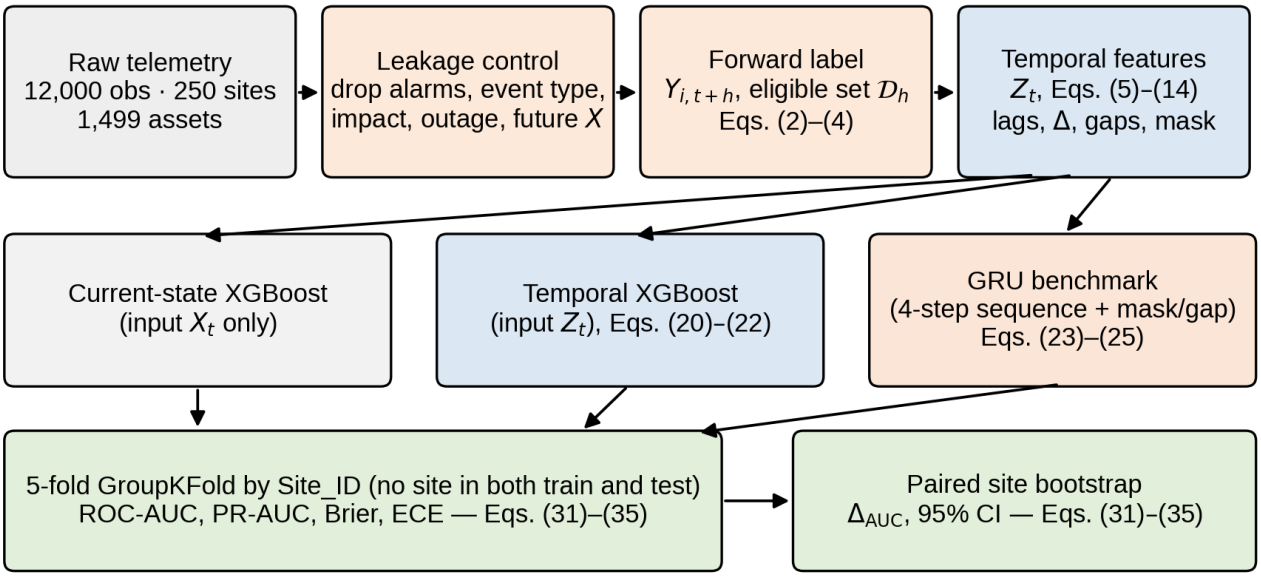


Fig. 3 End-to-end leakage-aware study design: raw telemetry, leakage control, forward labelling, temporal representation, the three compared learners, site-grouped cross-validation, and paired site-level bootstrap inference

6 Evaluation Protocol

For telecommunication-system generalization, the primary validation unit is the physical site. Five-fold GroupKFold is applied using Site_ID, so no observations from the same site are shared between training and test folds. This design addresses the dependence created by repeated observations from the same facility. Fold means and sample standard deviations are reported. Primary metrics are ROC-AUC, PR-AUC, and Brier score. Because future events are rare, PR-AUC is interpreted relative to the event prevalence rather than in isolation.

For a held-out set with positive set $\mathcal{P}$ (n_+ events) and negative set $\mathcal{N}$ (n_- non-events), the metrics are

$$\mathrm{AUC} = \frac{1}{n_+ n_-}\sum_{a\in\mathcal{P}}\sum_{b\in\mathcal{N}}\left[\mathbb{1}[\hat{p}_a > \hat{p}_b] + \frac{1}{2}\mathbb{1}[\hat{p}_a = \hat{p}_b]\right] \tag{26}$$

$$\mathrm{PR-AUC} \equiv \mathrm{AP} = \sum_q (R_q - R_{q-1})P_q \tag{27}$$

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(\hat{p}_n - Y_n)^2 \tag{28}$$

$$\mathrm{ECE} = \sum_{b=1}^{B_c}\frac{|\mathcal{B}_b|}{N}\left|\overline{Y}(\mathcal{B}_b) - \overline{p}(\mathcal{B}_b)\right| \tag{29}$$

$$\overline{A} = \frac{1}{5}\sum_{f=1}^{5} A_f, \quad s_A = \sqrt{\frac{1}{4}\sum_{f=1}^{5}\left(A_f - \overline{A}\right)^2} \tag{30}$$

Eq. (26) is the Mann–Whitney form of ROC-AUC [21]; in Eq. (27) R_q and P_q are recall and precision at the qth distinct score threshold, so that PR-AUC is reported as average precision and compared with π_h [23]; Eq. (28) is the Brier score [12]; in Eq. (29) $\mathcal{B}_b$ is the bth of B_c equal-width probability bins [22]; and in Eq. (30) A_f is a metric computed on held-out fold f. "Pooled OOF" values apply Eqs. (26)–(29) once to the concatenated out-of-fold predictions.

The paired model effect is calculated from the same held-out observations and assessed using site-level bootstrap resampling. For each bootstrap replicate, sites are sampled with replacement and the ROC-AUC difference between temporal and current-state predictions is recomputed. Three thousand bootstrap replicates are used for the primary 72-h comparison.

$$\Delta_{\mathrm{AUC}} = \mathrm{AUC}_{\mathrm{temporal}} - \mathrm{AUC}_{\mathrm{current}} \tag{31}$$

Let $\mathcal{S}$ be the set of 250 sites and $\mathcal{D}_s$ the held-out observations of site s. For replicate b, a multiset $\mathcal{S}^{*(b)}$ of $|\mathcal{S}|$ sites is drawn with replacement, and

$$\Delta_{\mathrm{AUC}}{}^{*(b)} = \mathrm{AUC}_{\mathrm{temporal}}\left(\mathcal{D}^{*(b)}\right) - \mathrm{AUC}_{\mathrm{current}}\left(\mathcal{D}^{*(b)}\right), \quad \mathcal{D}^{*(b)} = \bigcup_{s\in\mathcal{S}^{*(b)}} \mathcal{D}_s, \quad b = 1,\ldots,B \tag{32}$$

$$\mathrm{CI}_{95\%} = \left[Q_{0.025}\left(\left\{\Delta_{\mathrm{AUC}}{}^{*(b)}\right\}\right), Q_{0.975}\left(\left\{\Delta_{\mathrm{AUC}}{}^{*(b)}\right\}\right)\right], \quad \overline{\Delta_{\mathrm{AUC}}}^{*} = \frac{1}{B}\sum_{b=1}^{B}\Delta_{\mathrm{AUC}}{}^{*(b)} \tag{33}$$

where Q_α is the empirical α-quantile of the B replicates ($B = 3{,}000$ for the primary 72-h comparison). Because both models are scored on the same resampled observations, Eq. (32) is a paired statistic, and resampling whole sites preserves within-site dependence [20]. The effect is declared supported only when the interval in Eq. (33) excludes zero.

A key sensitivity analysis conditions on history availability. Because not every observation has three previous measurements, the temporal model is evaluated on progressively stricter subsets: at least one, at least two, and at least three prior observations. This test distinguishes genuine temporal signal from performance associated with irregular sequence availability.

The history-conditioned evaluation subsets and the rare-event lift statistic are

$$\mathcal{D}_h^{(k)} = \{(i,t) \in \mathcal{D}_h : n_i(t) \geq k\}, \quad k \in \{1,2,3\} \tag{34}$$

$$\Lambda_h = \frac{\mathrm{PR{-}AUC}_h}{\pi_h} \tag{35}$$

so that $\Lambda_h = 1$ corresponds to a no-skill ranking. The sample-retention fraction $\left|\mathcal{D}_h^{(k)}\right|/|\mathcal{D}_h|$ is reported alongside each subset.

Additional analyses. Four further analyses test whether the temporal gain reflects degradation dynamics or the observation process. In the mask-only ablation (E2), the lag values and change features are removed while the gap and availability channels are retained:

$$Z_t^{\mathrm{mask}} = [X_t, \Delta t_1, \Delta t_2, \Delta t_3, M_t] \tag{36}$$

Comparing the current-state, mask-only and temporal models yields the additive decomposition

$$\Delta_{\mathrm{AUC}} = (\mathrm{AUC}_{\mathrm{mask}} - \mathrm{AUC}_{\mathrm{current}}) + (\mathrm{AUC}_{\mathrm{temporal}} - \mathrm{AUC}_{\mathrm{mask}}) \equiv \Delta_{\mathrm{samp}} + \Delta_{\mathrm{dyn}} \tag{37}$$

where Δ_samp measures the ranking information carried by the observation pattern alone and Δ_dyn denotes the additional information in the lagged telemetry values. In the locked supplementary release, Δ_samp is formally tested with the paired site bootstrap of Eqs. (32)–(33), whereas Δ_dyn is reported descriptively because the locked primary temporal out-of-fold predictions were not regenerated in the supplementary re-estimation. In the long-window baselines (E1), each engineered variable is summarised over earlier observations of the asset within a trailing seven-day (168-h) window:

$$\mathcal{W}_i(t) = \{\tau : t - 168\,\mathrm{h} \leq \tau < t\}, \quad \mu_W = \frac{1}{n_W}\sum_{\tau\in\mathcal{W}_i(t)} x_\tau \tag{38}$$

$$s_W = \sqrt{\frac{1}{n_W - 1}\sum_{\tau\in\mathcal{W}_i(t)}\left(x_\tau - \mu_W\right)^2}, \quad x_W^{\min} = \min_{\tau\in\mathcal{W}_i(t)} x_\tau, \quad x_W^{\max} = \max_{\tau\in\mathcal{W}_i(t)} x_\tau \tag{39}$$

together with the window count $n_W = |\mathcal{W}_i(t)|$. The window uses only strictly earlier observations of the same asset, so no future or contemporaneous information enters the features; the statistics are computed for the principal electrical, thermal, cooling and network variables and evaluated with the same XGBoost configuration and five-fold site-grouped protocol. In the asset-class analysis (E3), with $a(i)$ the class of asset i, the class-conditional forward-event rate is

$$\mathcal{D}_{h,c} = \{(i,t) \in \mathcal{D}_h : a(i) = c\}, \quad \hat{\pi}_{h,c} = \frac{1}{|\mathcal{D}_{h,c}|}\sum_{(i,t)\in\mathcal{D}_{h,c}} Y_{i,t+h} \tag{40}$$

and its uncertainty is summarised by the Wilson score interval [24] with $n_c = |\mathcal{D}_{h,c}|$,

$$\mathrm{CI}_{95\%}(\hat{\pi}_{h,c}) = \frac{\hat{\pi}_{h,c} + \frac{z^2}{2n_c} \pm z\sqrt{\frac{\hat{\pi}_{h,c}\left(1-\hat{\pi}_{h,c}\right)}{n_c} + \frac{z^2}{4n_c{}^2}}}{1 + \frac{z^2}{n_c}}, \quad z = 1.96 \tag{41}$$

Homogeneity of the forward-event rate across the $C = 6$ asset classes is tested with Pearson's statistic on the $C \times 2$ table of observed counts O_{cy},

$$\chi^2 = \sum_{c=1}^{C}\sum_{y\in\{0,1\}}\frac{\left(O_{cy} - E_{cy}\right)^2}{E_{cy}}, \quad E_{cy} = \frac{n_c n_y}{N}, \quad \mathrm{df} = C - 1 = 5 \tag{42}$$

where n_y is the total number of observations with outcome y. Class-level results are reported descriptively; class-specific discrimination is not claimed because future-event counts per class range from 17 to 36.

6.1 Generalization stress tests

Two supplementary generalization tests are added without altering the locked primary experiment. First, a chronological split evaluates forward-period transfer. With τ the 70th percentile of reference timestamps in $\mathcal{D}_{72}$, training observations are purged so that no outcome window overlaps the test period:

$$\tau = Q_{0.7}(\{t\}), \quad \mathcal{D}_{\text{train}} = \{(i,t) : t + h \le \tau\}, \quad \mathcal{D}_{\text{test}} = \{(i,t) : t > \tau\} \tag{43}$$

Second, leave-one-asset-class-out (LOCO) training evaluates transfer across equipment classes: for each class c,

$$\mathcal{D}_{\text{train}}^{(c)} = \{(i,t) \in \mathcal{D}_h : a(i) \neq c\}, \quad \mathcal{D}_{\text{test}}^{(c)} = \{(i,t) \in \mathcal{D}_h : a(i) = c\} \tag{44}$$

Because each held-out set contains only 17–55 events and the site bootstrap cannot be applied to a single class-level split, the uncertainty of each held-out ROC-AUC is approximated with the Hanley–McNeil standard error [21],

$$\text{SE}_{\text{HM}}(A) = \sqrt{\frac{A(1-A) + (n_+ - 1)\left(Q_1 - A^2\right) + (n_- - 1)\left(Q_2 - A^2\right)}{n_+ n_-}}, \quad Q_1 = \frac{A}{2-A}, \quad Q_2 = \frac{2A^2}{1+A} \tag{45}$$

with n_+ and n_- the numbers of held-out events and non-events and a normal 95% interval $A \pm 1.96\,\text{SE}_{\text{HM}}$. Eq. (45) ignores within-site dependence and is therefore approximate and likely optimistic. Both tests are stress tests, not replacements for the site-grouped primary evaluation; results are reported in Section 7.8.

7 Results

7.1 72-hour primary experiment

The 72-h forward-risk experiment contains 9,991 eligible observations and 156 subsequent thermal/electrical events. The current-state XGBoost model performs near chance, with mean ROC-AUC 0.4869 ± 0.0310. The temporal XGBoost model reaches 0.5592 ± 0.0295, an absolute fold-mean improvement of approximately 0.0722. The pooled OOF ROC-AUC values are 0.4868 for the current-state model and 0.5587 for the temporal model. PR-AUC increases from 0.0162 ± 0.0016 to 0.0200 ± 0.0022, while Brier score remains similar at 0.01560 versus 0.01552.

The paired site-level bootstrap estimates a pooled ROC-AUC gain of 0.0741 with a 95% interval of 0.0235–0.1243. The interval excludes zero for this primary comparison, providing evidence that the temporal representation adds ranking information over the current-state baseline in the full eligible population.

Normalising PR-AUC by the 72-h prevalence (Eq. 35) gives a lift of 1.28 for temporal XGBoost, 1.21 for the GRU and 1.04 for the current-state baseline, i.e., the temporal representation concentrates future events in the upper risk ranks about 28% more densely than random ordering, whereas the current-state model is effectively at the no-skill level (Table 4, Fig. 4). Fig. 5 shows that the site-bootstrap interval for the full population lies entirely above zero.

Table 4 72-h forward-risk benchmark (5-fold site-grouped cross-validation; N = 9,991; 156 events)

Model	ROC-AUC mean ± SD	Pooled OOF ROC-AUC	PR-AUC mean ± SD	PR lift Λ_{72}[a]	Brier mean ± SD	Pooled ECE[b]
Current-state XGBoost	0.4869 ± 0.0310	0.4868	0.0162 ± 0.0016	1.04	0.01560 ± 0.00087	0.0020
Temporal XGBoost	**0.5592 ± 0.0295**	**0.5587**	**0.0200 ± 0.0022**	**1.28**	**0.01552 ± 0.00090**	**0.0042**
GRU	0.5281 ± 0.0334	0.5263	0.0188 ± 0.0036	1.21	0.01542 ± 0.00088	—

[a] Λ_{72} = mean PR-AUC / π_{72} with π_{72} = 156/9,991 = 0.0156 (Eq. 35) [b] Not computed for the GRU. Bold row: principal temporal learner; SD is the sample standard deviation over five site-grouped folds (Eq. 30)

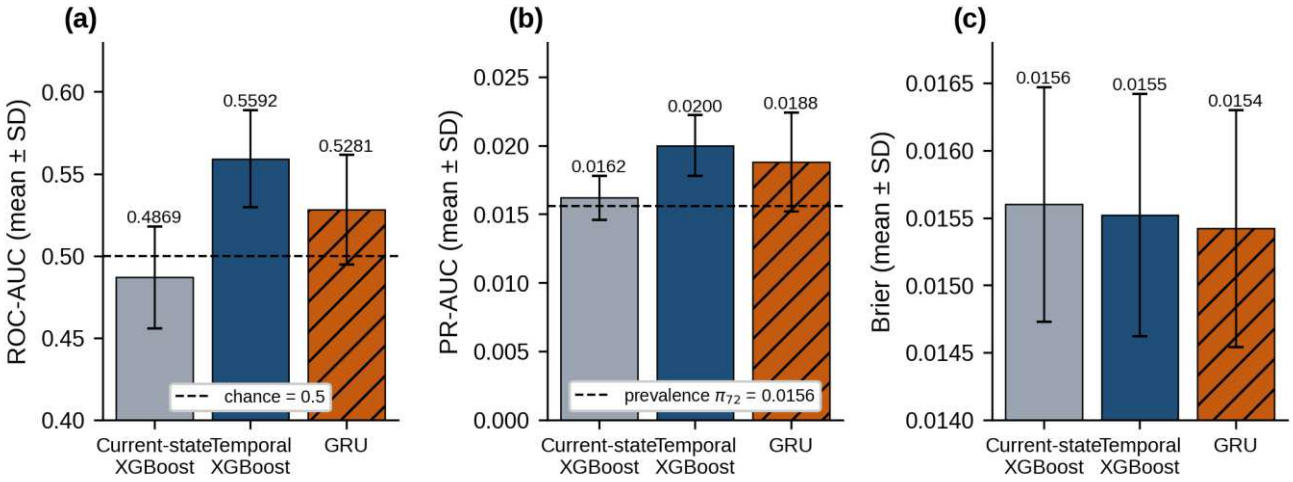


Fig. 4 72-h thermal/electrical forward-risk benchmark: (a) ROC-AUC with chance reference, (b) PR-AUC with prevalence reference π_{72}, (c) Brier score. Error bars show fold-level sample SD

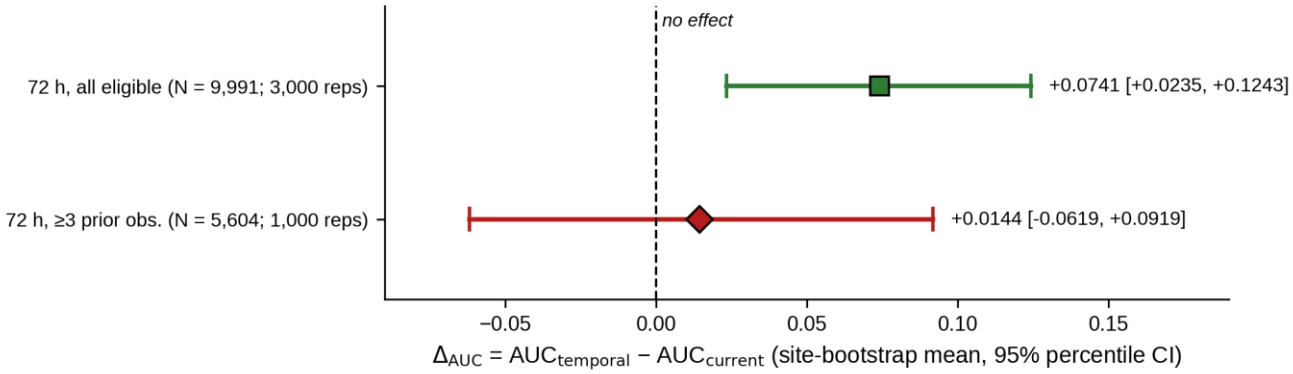


Fig. 5 Paired site-level bootstrap estimates of Δ_AUC (Eqs. 32–33) at 72 h for the full eligible population (3,000 replicates) and the ≥3-prior-observation subset (1,000 replicates). Markers: bootstrap mean; bars: 95% percentile interval; dashed line: no effect. Only the full-population interval excludes zero

7.2 Horizon sensitivity

Temporal benefit is not monotonic with prediction horizon. At 24 h, the temporal representation does not improve pooled AUC over the current-state model. At 48 h, the pooled AUC increases modestly. The strongest observed difference occurs at 72 h. At 96 h the gain becomes smaller, and at 168 h (7 d) the temporal model falls below the current-state baseline. This pattern indicates a limited useful horizon under the available sampling structure rather than a universal long-range forecasting capability.

Table 5 Horizon sensitivity of the forward-risk task

Horizon h	Prevalence π_h	Current-state ROC-AUC	Temporal ROC-AUC	Δ_AUC(h)	Temporal PR-AUC	Lift Λ_h[a]
24 h	0.50%	0.5484	0.5382	−0.0101	0.0174	3.48
48 h	1.06%	0.5145	0.5324	+0.0179	0.0139	1.31
72 h	**1.56%**	**0.4869**	**0.5592**	**+0.0722**	**0.0200**	**1.28**
96 h	2.12%	0.4968	0.5184	+0.0216	0.0240	1.13
168 h	3.38%	0.5117	0.4598	−0.0519	0.0308	0.91

[a]Computed from the reported PR-AUC and rounded prevalence (Eq. 35); the 24-h value is therefore approximate. Bold row: primary horizon

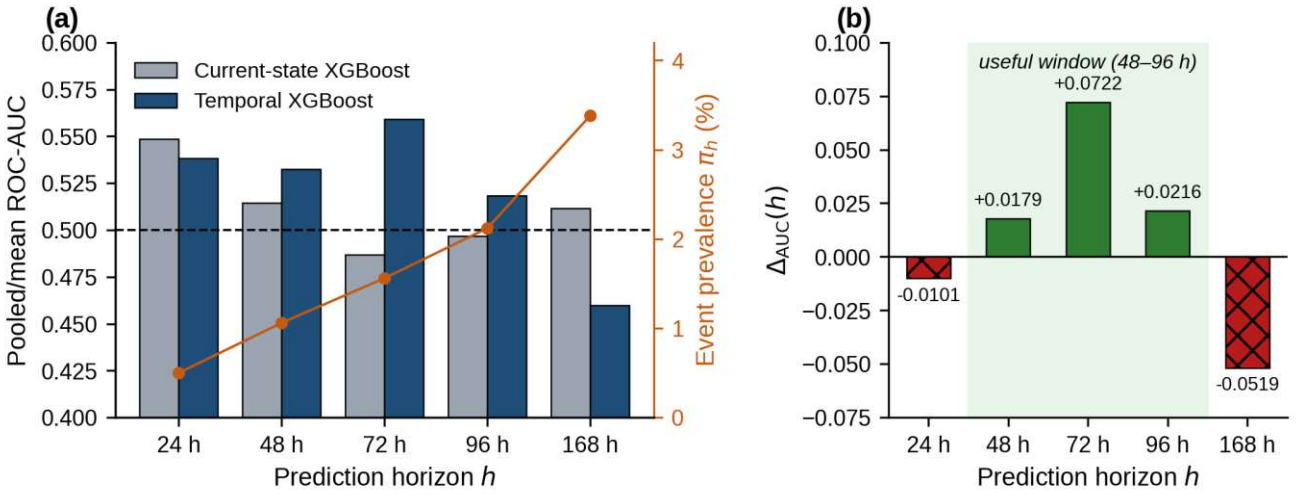


Fig. 6 Horizon sensitivity. (a) Current-state and temporal ROC-AUC with forward-event prevalence π_h (right axis). (b) Temporal gain Δ_AUC(h); the shaded region marks the 48–96 h window with positive gain

Two further observations follow from Table 5. First, prevalence rises almost sevenfold from 24 h to 168 h, so the horizon trend is not driven by a shrinking event base. Second, the PR-AUC lift declines monotonically with horizon (Fig. 7) and falls below one at 168 h. The comparatively high 24-h lift coexists with a sub-chance ROC-AUC gain and rests on very few events at 0.50% prevalence, so it is not interpreted as evidence of short-horizon skill.

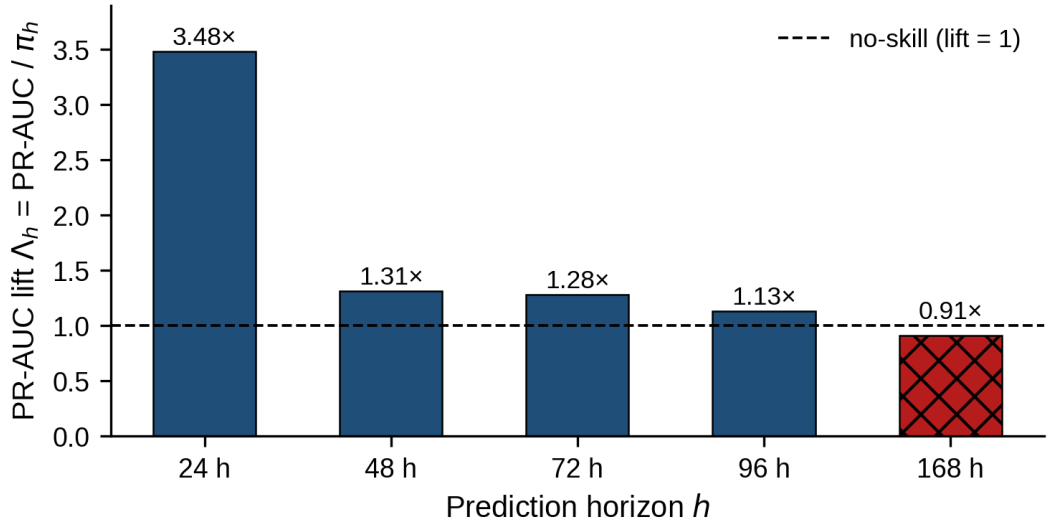


Fig. 7 PR-AUC lift of the temporal model over the forward-event prevalence (Eq. 35) by horizon. The dashed line marks no-skill ranking

7.3 History-availability sensitivity

The main qualification is the sparsity of temporal history. In the full eligible population, the temporal model uses a mixture of complete and incomplete histories with explicit missing-lag indicators. Restricting the evaluation to observations with at least one prior measurement reduces the pooled gain to 0.0448. With at least two prior observations the gain is effectively zero (0.0000). With at least three prior observations, the pooled gain is 0.0144, and a 1,000-replicate paired site bootstrap gives a 95% interval of approximately −0.0618 to 0.0919. The early-warning gain is therefore not robust to strict history-completeness restrictions.

Table 6 also shows that the forward-event rate remains between 1.45% and 1.56% across all subsets, so the attenuation of the gain cannot be attributed to a change in class balance; it coincides with the progressive removal of observations whose history is partly imputed (Fig. 8).

Table 6 History-availability sensitivity at 72 h (pooled out-of-fold ROC-AUC)

History requirement	N	Retained[a]	Events	Event rate	Current AUC	Temporal AUC	Δ_AUC^(k)[b]

History requirement	N	Retained[a]	Events	Event rate	Current AUC	Temporal AUC	Δ_AUC^(k)[b]
All eligible	9,991	100.0%	156	1.56%	0.4868	0.5587	+0.0718
≥1 prior observation	8,498	85.1%	131	1.54%	0.4850	0.5298	+0.0448
≥2 prior observations	7,032	70.4%	102	1.45%	0.5202	0.5202	+0.0000
≥3 prior observations	5,604	56.1%	87	1.55%	0.4798	0.4942	+0.0144

[a] $|\mathcal{D}_{72}^{\wedge}(k)| / |\mathcal{D}_{72}|$ (Eq. 34) [b] For ≥3 prior observations, the 1,000-replicate site-bootstrap 95% CI is −0.0618 to 0.0919

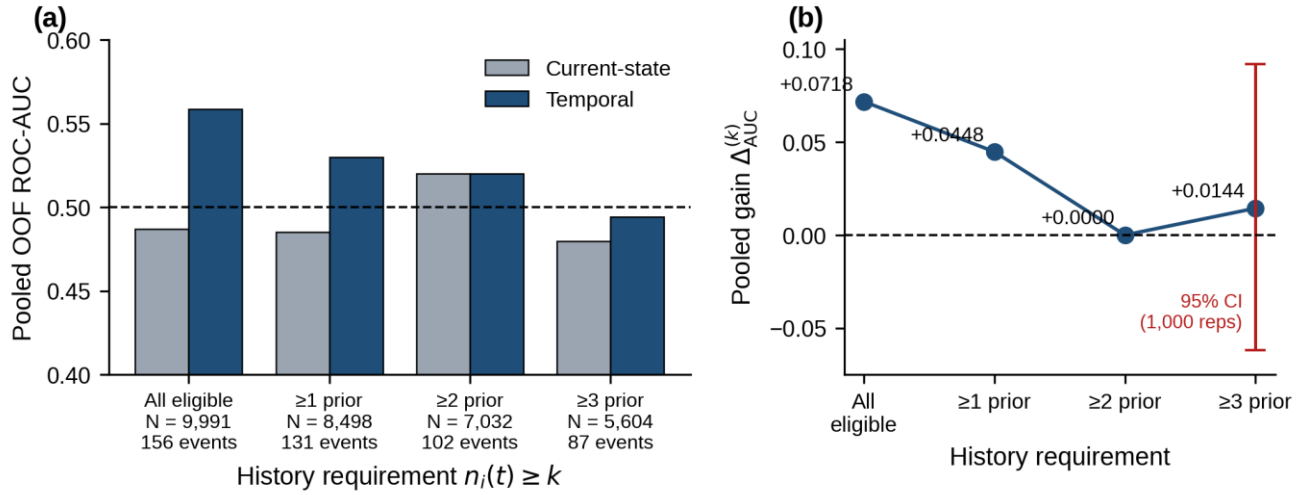


Fig. 8 History-availability sensitivity. (a) Pooled OOF ROC-AUC of both models by minimum number of prior observations, with subset size and event count. (b) Pooled temporal gain, with the 95% site-bootstrap interval for the ≥3-prior subset

7.4 Recurrent benchmark

The GRU benchmark reaches 0.5281 ± 0.0334 mean ROC-AUC and pooled OOF ROC-AUC 0.5263. This is below the temporal XGBoost result. The finding should not be interpreted as evidence that recurrent neural networks are unsuitable in general; rather, within this dataset, sequence length, sparse sampling, small event counts, and the fixed four-step input may favour a simpler lag-derived representation.

7.5 Sampling-pattern versus dynamic information

Table 7 reports the mask-only ablation at the 72-h horizon. The mask-only model receives the same current-state variables as the primary baseline but removes lagged telemetry values and change features, retaining only the lag-availability indicators M_{t,k} and elapsed-time variables Δt_k. It achieves pooled OOF ROC-AUC 0.5021, only slightly above the current-state baseline (0.4868). A paired 3,000-resample site bootstrap gives a mask-minus-current AUC difference of 0.0156 with a 95% confidence interval of −0.0085 to 0.0395, which includes zero. Thus, the observation pattern alone does not provide statistically supported aggregate discrimination. The locked primary temporal model remains higher at pooled OOF ROC-AUC 0.5587; its descriptive gap over mask-only is 0.0566. Because the locked temporal OOF predictions were not regenerated in this supplementary re-estimation, the temporal-minus-mask quantity is reported descriptively rather than assigned a new paired confidence interval.

On pooled out-of-fold values, the total gain therefore decomposes (Eq. 37) into $\Delta_{\text{samp}} = 0.0153$ and $\Delta_{\text{dyn}} = 0.0566$, i.e., approximately 21% and 79% of the pooled improvement of 0.0719 (Fig. 9b). Only the first component carries a paired interval, and that interval includes zero.

Table 7 Mask-only ablation and gain decomposition at 72 h (5-fold site-grouped cross-validation)

Model / component	Pooled ROC-AUC	ROC-AUC mean ± SD	Pooled PR-AUC	Interpretation
Current-state XGBoost	0.4868	0.4869 ± 0.0310	0.0162	Non-temporal control
Mask-only XGBoost (Eq. 36)	0.5021	0.5020 ± 0.0242	0.0187	Sampling-pattern control
Three-lag temporal XGBoost (locked primary)	**0.5587**	**0.5592 ± 0.0295**	**0.0200**	**Primary temporal representation**
Δ_samp = mask − current[a]	+0.0156	—	—	95% CI [−0.0085, 0.0395]; includes zero
Δ_dyn = temporal − mask[b]	+0.0566	—	—	Descriptive; no paired CI

[a]Site-bootstrap mean of the paired difference, 3,000 replicates (Eqs. 32–33); the pooled point difference is 0.0153 [b]Difference of pooled OOF ROC-AUC; the locked temporal OOF predictions were not regenerated, so no paired interval is reported

7.6 Long-window rolling baselines

Table 8 adds a seven-day (168-h) trailing rolling-statistics baseline. For each reference observation, the rolling features are calculated only from strictly earlier observations of the same asset, using mean, standard deviation, minimum, maximum, and observation count for the principal electrical, thermal, cooling and network variables. This provides a longer-history control without using future data. The seven-day rolling model achieves pooled OOF ROC-AUC 0.4745 and mean 0.4764 ± 0.0570, below the current-state baseline and well below the locked three-lag temporal model. Its paired site-bootstrap difference versus current state is −0.0123 (95% CI −0.0517 to 0.0239), which includes zero. The result therefore does not support the interpretation that simply extending the history window with aggregate statistics recovers the observed three-lag temporal benefit.

Table 8 Long-window rolling baselines at 72 h

Representation / learner	Pooled ROC-AUC	ROC-AUC mean ± SD	Pooled PR-AUC	Δ vs current [95% CI]
Current state, XGBoost	0.4868	0.4869 ± 0.0310	0.0162	Reference
7-day rolling statistics, XGBoost (Eqs. 38–39)	0.4745	0.4764 ± 0.0570	0.0154	−0.0123 [−0.0517, 0.0239][a]
Three-lag temporal, XGBoost (locked primary)	**0.5587**	**0.5592 ± 0.0295**	**0.0200**	**+0.0741 [0.0235, 0.1243][b]**

[a]Paired site bootstrap, 3,000 replicates [b]Locked primary 72-h result (Section 7.1): bootstrap mean 0.0741; the pooled point difference is 0.0718

7.7 Asset-class analysis

Table 9 disaggregates the 72-h forward population by asset class. The analysis reports event counts and rates rather than class-specific model-ranking claims because future voltage anomalies are sparse and the number of positive outcomes varies substantially across classes. This is an outcome-stratification analysis, not a replacement for the primary site-grouped model evaluation.

Future-event rates range from 0.98% (HVAC) to 2.03% (Generator), but Pearson's test does not reject homogeneity across the six classes ($\chi^2(5) = 8.19$, $p = 0.146$; Eq. 42), and every class interval overlaps the pooled rate of 1.56%. The class ordering in Fig. 10b should therefore be read as indicative rather than established.

Table 9 Asset-class analysis of the 72-h forward-risk task

Asset class	72-h eligible N	Future TE events	Future event rate [95% CI][a]	Recorded overheating (rate)[b]	Recorded voltage anomalies
HVAC	2,049	20	0.98% [0.63, 1.50]	103 (4.21%)	8
Generator	1,775	36	2.03% [1.47, 2.79]	83 (3.86%)	10
UPS/Rectifier	1,802	33	1.83% [1.31, 2.56]	80 (3.75%)	13
DC Rectifier	1,650	27	1.64% [1.13, 2.37]	74 (3.67%)	10
RF Power Amp	1,508	23	1.53% [1.02, 2.28]	79 (4.40%)	0
Transfer Switch	1,207	17	1.41% [0.88, 2.24]	52 (3.57%)	6
All classes	**9,991**	**156**	**1.56% [1.34, 1.82]**	**471 (3.93%)**	**47**

[a]Wilson score interval (Eq. 41); homogeneity across classes $\chi^2(5) = 8.19$, $p = 0.146$ (Eq. 42) [b]Recorded counts and rates use all 12,000 observations; future-event quantities use the 72-h eligible population

Fig. 9 provides the combined sampling-pattern and longer-history controls, while Fig. 10 shows where the thermal/electrical event burden is concentrated across infrastructure classes. The two controls provide complementary evidence: mask/gap variables alone do not reproduce the temporal result, and a simple seven-day rolling representation does not recover the gain.

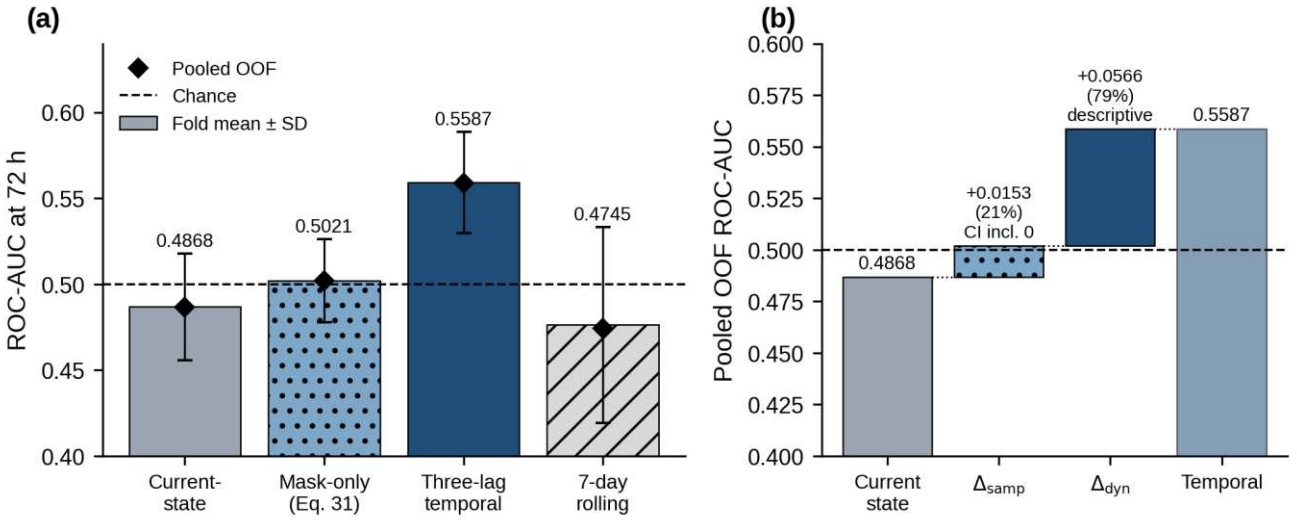


Fig. 9 Sampling-pattern and longer-history controls at 72 h. (a) Fold-mean ROC-AUC ± sample SD (bars) and pooled OOF ROC-AUC (diamonds) for the current-state, mask-only, three-lag temporal and seven-day rolling models; the temporal result is the locked primary benchmark. (b) Additive decomposition of the pooled gain into Δ_samp and Δ_dyn (Eq. 37); dashed lines mark chance

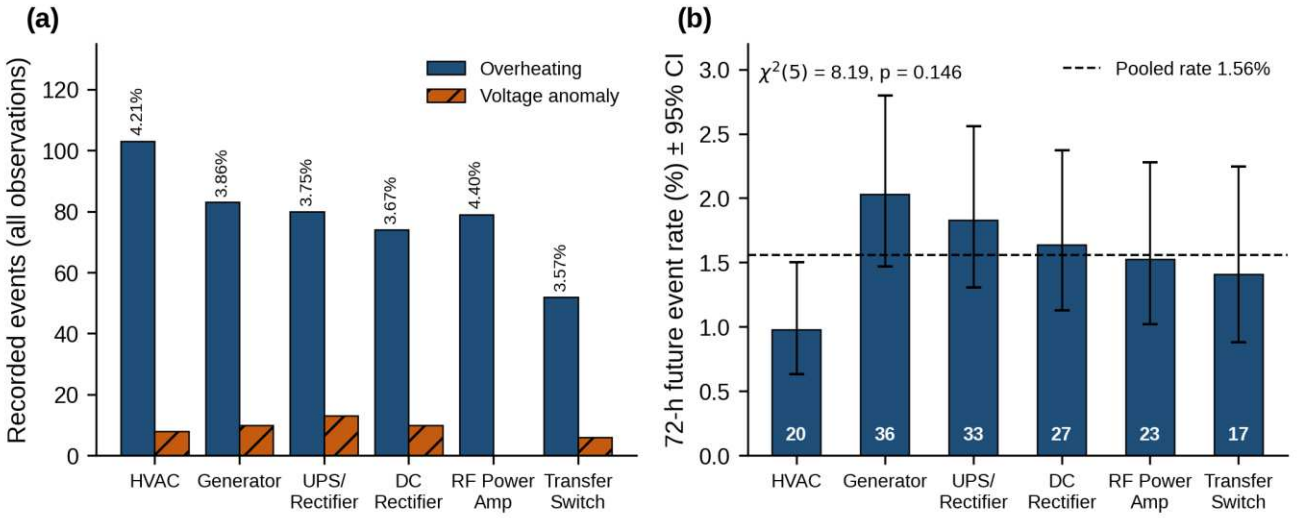


Fig. 10 Thermal/electrical events by asset class. (a) Recorded overheating and voltage-anomaly counts in all 12,000 observations, with class overheating rates. (b) 72-h future-event rate with Wilson 95% intervals (Eq. 41) and event counts; the dashed line marks the pooled rate, and the homogeneity test is $\chi^2(5) = 8.19$, $p = 0.146$

7.8 Reliability-gating ablation and generalization stress tests

The proposed reliability-aware framework was tested with an exploratory learned gate that combines history availability, elapsed observation gaps, temporal consistency and physics-informed stress indicators. The gated variant (Eq. 19) achieves pooled ROC-AUC 0.5220 and PR-AUC 0.0162, compared with 0.5611 and 0.0180 for the re-estimated three-lag temporal learner used in this supplementary experiment. Its Brier score is slightly lower (0.01549 versus 0.01553), indicating a modest probabilistic improvement without a corresponding ranking improvement. A paired site bootstrap gives gated-minus-current-state ΔAUC = 0.0349 [0.0109, 0.0583], but gated-minus-temporal ΔAUC = −0.0392 [−0.0657, −0.0115]. The gate is therefore retained as an evidence-quality/deployment concept rather than claimed as a superior predictive learner. The locked primary temporal result remains unchanged.

Table 10 Exploratory reliability-gating ablation at 72 h (5-fold site-grouped cross-validation)

Model / comparison	Pooled ROC-AUC	Pooled PR-AUC	Brier score	Paired ΔAUC [95% CI][a]

Model / comparison	Pooled ROC-AUC	Pooled PR-AUC	Brier score	Paired ΔAUC [95% CI][a]
Three-lag temporal XGBoost (re-estimated)[b]	0.5611	0.0180	0.01553	—
Reliability-gated variant (Eq. 19)	0.5220	0.0162	0.01549	—
Gated − current state	—	—	—	+0.0349 [0.0109, 0.0583]
Gated − temporal	—	—	—	−0.0392 [−0.0657, −0.0115]

[a]Paired site bootstrap (Eqs. 32–33) [b]Re-estimated within the gating experiment; the locked primary result (0.5587) is unchanged

Table 10 shows that the gate improves on the current-state baseline but is significantly worse than the temporal learner in ranking, while the Brier-score difference (0.00004) is negligible.

Chronological and asset-class stress tests provide a second boundary (Table 11, Fig. 11). Under the chronological split of Eq. (43), temporal XGBoost obtains ROC-AUC 0.5015 (Hanley–McNeil 95% CI 0.424–0.579) and PR-AUC 0.0177 on 2,999 held-out observations with 55 events. Leave-one-asset-class-out ROC-AUC ranges from 0.4956 (DC Rectifier) to 0.6852 (RF Power Amp). Only the RF Power Amp interval excludes 0.5 ($z = 2.98$, $p = 0.003$; Bonferroni-adjusted over seven tests, $p = 0.020$), and every other interval includes chance. Because the held-out event counts are small and performance varies materially by class, these tests do not support a claim of universal transfer; site-grouped generalization is the strongest supported validation regime for the present dataset, while temporal-period and cross-class transfer require denser, multi-environment data.

Table 11 Generalization stress tests for temporal XGBoost at 72 h

Held-out set	N	Events	ROC-AUC	95% CI (Hanley–McNeil)[a]
Site-grouped 5-fold (primary, pooled OOF)	9,991	156	0.5587	Δ vs current: [0.0235, 0.1243][b]
Chronological 70/30, 72-h purge (Eq. 43)	2,999	55	0.5015	[0.424, 0.579]
LOCO: DC Rectifier	1,650	27	0.4956	[0.386, 0.605]
LOCO: Generator	1,775	36	0.4972	[0.402, 0.592]
LOCO: HVAC	2,049	20	0.6042	[0.472, 0.736]
LOCO: RF Power Amp	**1,508**	**23**	**0.6852**	**[0.564, 0.807]**
LOCO: Transfer Switch	1,207	17	0.5660	[0.424, 0.709]
LOCO: UPS/Rectifier	1,802	33	0.5389	[0.437, 0.640]

[a]Eq. (45); ignores within-site dependence and is therefore approximate. LOCO held-out N and events are the 72-h eligible class counts of Table 9; chronological training N = 6,530 [b]Paired site-bootstrap interval of the temporal-minus-current difference (Section 7.1)

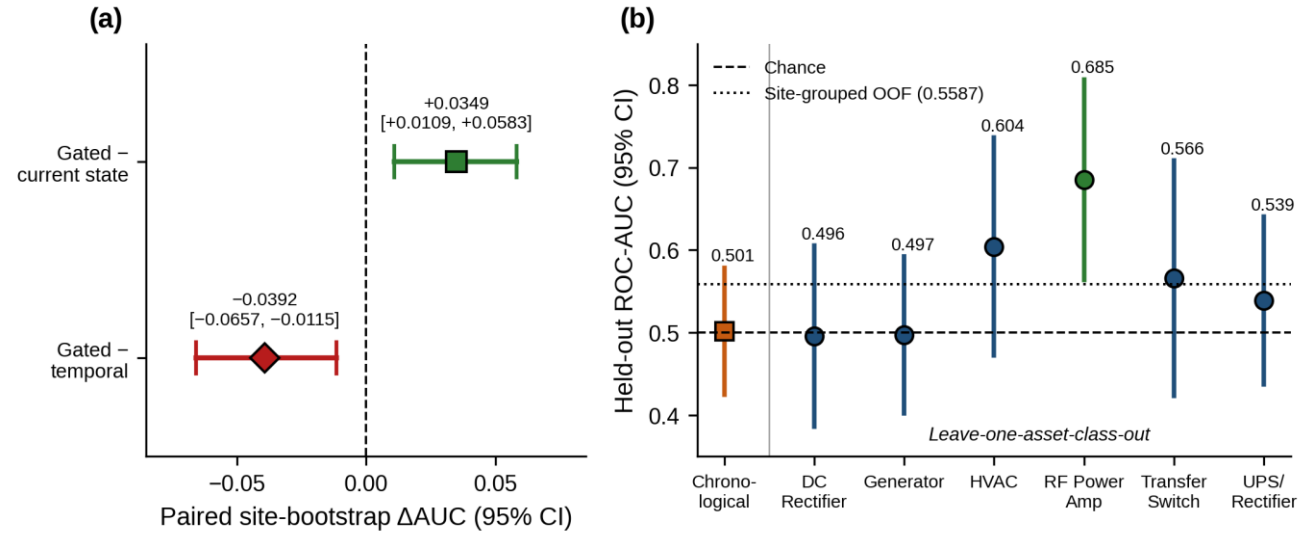


Fig. 11 Reliability-gating ablation and generalization stress tests. (a) Paired site-bootstrap ΔAUC of the gated variant against the current-state and temporal learners (95% CIs; dashed line: no difference). (b) Held-out ROC-AUC with Hanley–McNeil 95% intervals for the chronological split and each leave-one-asset-class-out split; dashed line: chance, dotted line: primary site-grouped pooled OOF ROC-AUC

8 Discussion

Table 12 consolidates the evidence for the six research questions.

Table 12 Evidence summary for RQ1–RQ6

RQ	Question	Evidence	Verdict
RQ1	Does temporal history improve future event discrimination?	Δ_AUC = +0.0741, 95% CI [0.0235, 0.1243] (full 72-h population)	Supported, conditionally
RQ2	At which lead times?	Δ_AUC > 0 at 48, 72, 96 h; < 0 at 24 h and 168 h	Useful window 48–96 h
RQ3	Does the gain persist with complete history?	≥2 prior: 0.0000; ≥3 prior: +0.0144, CI [−0.0618, 0.0919]	Not supported
RQ4	GRU vs tree-based temporal representation?	GRU 0.5281 ± 0.0334 vs temporal XGBoost 0.5592 ± 0.0295	Tree-based preferred here
RQ5	Appropriate deployment claim?	Modest AUC, lift 1.28 at 72 h, history-dependent gain; chronological AUC 0.5015; LOCO 0.4956–0.6852	Prioritisation aid with human review
RQ6	Is the gain driven by the sampling pattern rather than dynamics?	Mask-only AUC 0.5021 vs current 0.4868; paired site bootstrap +0.0156, 95% CI [−0.0085, 0.0395]. Seven-day rolling pooled AUC 0.4745.	Sampling-pattern alone does not reproduce temporal gain; longer-window aggregation does not recover it. Gain remains history-dependent.

The central finding is conditional and reliability-aware. In the full set of observations with complete 72-h follow-up, adding recent multivariate history improves future thermal/electrical ranking relative to the current-state baseline. The improvement is statistically supported by a site-level paired bootstrap. However, the framework does not treat temporal history as universally reliable: history-completeness sensitivity, the negative reliability-gating ablation and chronological stress test all indicate that the temporal signal is bounded by the sampling regime and operational environment. The principal contribution is therefore not a claim that a more complex temporal learner always wins, but a disciplined framework for determining when temporal evidence adds information and when the evidence is too weak to support a stronger claim.

At the same time, the strict history-availability analysis limits the strength of that conclusion. When at least two prior observations are required, the observed advantage disappears. This suggests that the full-population temporal gain may partly depend on the irregular observation process and the information carried by the presence or absence of historical measurements. In other words, temporal information is useful in the available data regime, but the current dataset is not sufficiently dense to demonstrate a robust sequence model that works independently of sampling structure.

The mask-only result is especially informative for causal interpretation of the temporal gain. Its pooled ROC-AUC of 0.5021 is close to the current-state control (0.4868), and the paired site-bootstrap interval for mask-minus-current includes zero. Therefore, the presence or absence of earlier measurements and their elapsed-time gaps do not, by themselves, explain the full-population temporal result. The seven-day rolling baseline also fails to recover the primary gain. Taken together, these controls are consistent with the interpretation that lagged signal changes contribute to the observed 72-h improvement, while the history-completeness sensitivity shows that this contribution is not robust when only well-populated histories are retained.

Sections 7.5–7.7 test this interpretation directly: the mask-only control isolates the observation-pattern contribution, the seven-day rolling baseline tests whether a longer but simpler history recovers the signal, and the asset-class analysis identifies where the future event burden is concentrated.

Asset stratification adds an operational perspective. Recorded overheating is most frequent in HVAC (103 observations) and RF Power Amplifier (79 observations), with overheating rates of 4.21% and 4.40%, respectively. In the 72-h forward population, however, future thermal/electrical event counts are highest for Generators (36) and UPS/Rectifiers (33). This difference illustrates why event prevalence and model discrimination should not be conflated and why class-specific findings require larger event counts before being used for class-specific automation.

The horizon sensitivity provides a second boundary. The benefit is concentrated around 48–96 h and is largest at 72 h. At seven days the temporal model no longer improves performance. This is consistent with the median site observation gap of

roughly 23.75 h in the underlying study and with the general difficulty of learning long-horizon failure dynamics from sparse observations.

Operationally, the results support a cautious reliability-aware early-risk workflow. A 72-h temporal risk score may be used as a prioritisation signal when recent historical measurements are available, but it should be accompanied by a history-completeness indicator, elapsed-gap information, temporal-consistency status and human engineering review. Where the temporal evidence is weak or stale, the current-state model should remain the reference rather than forcing a temporal interpretation. The result is better interpreted as early risk identification with an evidence-quality qualifier than as a validated remaining-useful-life estimator or deterministic failure-time predictor.

The study also contributes to telecommunication-system management by connecting equipment condition with network operations. Thermal/electrical risk at a cellular site has consequences for service continuity, maintenance dispatch, energy management, and network availability. This systems-level view is consistent with the reliability and management orientation of *Telecommunication Systems*, which has published work on reliable communication systems [16], machine-learning-assisted resource monitoring in 5G systems [17], and learning-based methods in wireless communications [18, 19].

9 Operational Deployment Architecture

A practical deployment architecture comprises: (1) telemetry ingestion; (2) timestamp alignment and data-quality screening; (3) current-state and temporal feature generation; (4) forward-risk inference; (5) calibration and history-completeness checks; and (6) engineer-facing decision support. The system should not automatically initiate costly maintenance solely from a model score. Instead, a high-risk result should trigger corroboration using current physical measurements and alarm provenance.

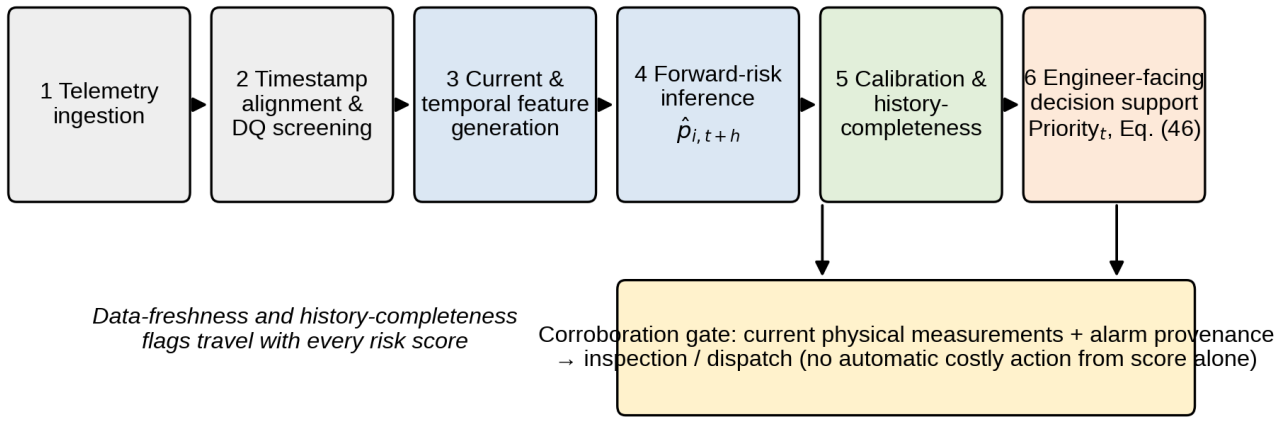


Fig. 12 Operational deployment architecture. Risk scores are always accompanied by data-freshness and history-completeness flags and pass a corroboration gate before any costly field action

$$\text{Priority}_t = w_r R_t + w_c C_t + w_s S_t, \quad w_\bullet \geq 0, \; w_r + w_c + w_s = 1 \tag{46}$$

where R_t is the calibrated model risk, C_t is asset/site criticality, S_t is service-impact importance (each normalised to [0, 1]), and the weights are deployment-policy parameters rather than learned quantities. The formula is illustrative and should be calibrated to the operator's cost of false alarms, missed events, and inspection workload.

10 Limitations and Threats to Validity

First, the source dataset covers one operational environment and approximately ten weeks. The added chronological 70/30 stress test is close to chance, and leave-one-asset-class-out performance is heterogeneous; therefore the study cannot establish universal generalization across operators, regions, weather regimes, seasons, equipment manufacturers or future operating periods. Second, future voltage anomalies are rare relative to overheating events, limiting event-specific inference. Third, observations are irregular and the median site-level gap is about 23.75 h; this restricts the ability to reconstruct continuous degradation trajectories. Fourth, the target represents a subsequent observed abnormal state, not physical component failure time or remaining useful life. Fifth, the temporal lag representation uses imputation and explicit history-availability indicators, and the mask-only, seven-day rolling and reliability-gating controls show that temporal

advantage cannot be interpreted independently of the observation process. Sixth, the study does not identify causality: a temporal feature that improves ranking does not establish a physical failure mechanism. Seventh, the study does not claim that the exploratory reliability gate is a superior learner; its value is as a deployment evidence-quality concept. Finally, the model is not a complete digital twin or first-principles thermal simulator, and the title's term early risk is operationally defined as forward-risk assessment at specified horizons.

11 Reproducibility and Data Availability

The raw telemetry workbook is proprietary and is not redistributed. The reproducibility package accompanying this manuscript contains the analysis configuration, feature definitions, fold-level summary results, horizon sensitivity results, the mask-only and seven-day rolling control outputs, asset-class event summaries, bootstrap outputs, figures, manuscript source, and scripts describing the evaluation protocol. Authorized analysts can reproduce the computations by placing the confidential workbook at the documented data path. The supplementary control files include maskonly_folds.csv, bootstrap_mask-current.csv, 7day_rolling_folds.csv, bootstrap_rolling7d_current.csv, and asset_class_current_rolling.csv. No proprietary row-level predictions are required for these aggregate checks.

All preprocessing, temporal feature generation, model fitting, and calibration operations must be performed within each training fold. Site_ID is used as the grouping variable only. The package records the exact random seed, estimator settings, horizon definitions, and history-sensitivity protocol used for the reported results.

12 Conclusion

This study examined whether multivariate temporal history improves early identification of thermal and electrical abnormal states in cellular network infrastructure. Using real telemetry from 12,000 observations, 250 sites, and 1,499 assets, the 72-h temporal XGBoost model achieved mean ROC-AUC 0.5592 ± 0.0295 compared with 0.4869 ± 0.0310 for the current-state baseline. The paired site-level bootstrap estimated an OOF AUC gain of 0.0741 (95% CI 0.0235–0.1243). The study's methodological contribution is a reliability-aware temporal learning framework that separates temporal signal from sampling-pattern information and explicitly qualifies temporal evidence using history completeness, elapsed observation gaps and temporal consistency. Mask-only controls achieve pooled OOF ROC-AUC 0.5021, the seven-day rolling baseline 0.4745, and the exploratory learned reliability gate does not outperform the temporal learner. Chronological and leave-one-asset-class-out stress tests further show that generalization beyond site-disjoint validation remains limited and heterogeneous. The evidence therefore supports a bounded 72-h early-risk prioritisation capability rather than a universal predictive-maintenance claim. Future work should use denser telemetry, multi-operator and multi-season validation, larger event counts, explicit missing-sequence modelling, genuine failure-time or lead-time labels, and independently evaluated selective-prediction mechanisms. The study is framed as a telecommunication-system reliability and management problem, with temporal learning used as a component of a practical monitoring and decision-support loop.

References

[1] Correia, L., Goos, J.-C., Klein, P., Bäck, T., & Kononova, A. V. (2024). Online model-based anomaly detection in multivariate time series: Taxonomy, survey, research challenges and future directions. *Engineering Applications of Artificial Intelligence, 138*, 109323. https://doi.org/10.1016/j.engappai.2024.109323

[2] Wang, F., Jiang, Y., Zhang, R., Wei, A., Xie, J., & Pang, X. (2025). A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors, 25*(1), 190. https://doi.org/10.3390/s25010190

[3] Jia, X., Xun, P., Peng, W., Zhao, B., Li, H., & Shen, C. (2025). Deep anomaly detection for time series: A survey. *Computer Science Review, 58*, 100787. https://doi.org/10.1016/j.cosrev.2025.100787

[4] Syed, M. A. B., Hasan, M. R., Chowdhury, N. I., Rahman, M. H., & Ahmed, I. (2025). A systematic review of time series algorithms and analytics in predictive maintenance. *Decision Analytics Journal, 15*, 100573. https://doi.org/10.1016/j.dajour.2025.100573

[5] Ahmed, S., Terefe Debella, T., & Hailemariam, D. (2024). Machine learning for base transceiver stations power failure prediction: A multivariate approach. *e-Prime – Advances in Electrical Engineering, Electronics and Energy, 10*, 100814. https://doi.org/10.1016/j.prime.2024.100814

[6] Anomaly detection and classification in cellular networks using automatic labeling technique for applying supervised learning. (2018). *Procedia Computer Science, 140*, 186–195. https://doi.org/10.1016/j.procs.2018.10.328

[7] *Predicting and mitigating failure of telecommunications network equipment* (U.S. Patent No. 12,563,421 B2). (2026). U.S. Patent and Trademark Office.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[9] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[13] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. https://doi.org/10.1145/1102351.1102430

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.

[16] Rak, J., Pickavet, M., Trivedi, K. S., Alonso Lopez, J., Koster, A. M. C. A., Sterbenz, J. P. G., Çetinkaya, E. K., Gomes, T., Gunkel, M., Walkowiak, K., et al. (2015). Future research directions in design of reliable communication systems. *Telecommunication Systems, 60*, 423–450. https://doi.org/10.1007/s11235-015-9987-7

[17] Conţu, C., Borcoci, E., Vochin, M.-C., et al. (2025). Enabling machine learning-assisted resource monitoring for network slice creation and management in OSM: Design, implementation, and validation. *Telecommunication Systems, 88*, 109. https://doi.org/10.1007/s11235-025-01342-7

[18] Tsipi, L., Karavolos, M., Papaioannou, G., et al. (2024). Machine learning-based methods for MCS prediction in 5G networks. *Telecommunication Systems, 86*, 705–728. https://doi.org/10.1007/s11235-024-01158-x

[19] ElSabaa, A., Guéniat, F., & Wu, W. (2026). Data-driven single gateway location estimation using improved channel model and LSTM. *Telecommunication Systems, 89*, 59. https://doi.org/10.1007/s11235-026-01429-9

[20] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593

[21] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[22] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR.

[23] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[24] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

Statements and Declarations

Funding The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests. Financial interests: the author declares no financial interests.

Author contributions J.A. conceived the study, curated and analyzed the data, developed the methodology, performed the experiments, and wrote and revised the manuscript.

Data availability The datasets generated during and/or analysed during the current study are not publicly available due to the fact that operational data is proprietary and subject to organizational confidentiality and data-governance restrictions but are available from the corresponding author on reasonable request.

Code availability The analysis scripts, configuration files, derived summary outputs, and figure-generation procedures are provided in the accompanying reproducibility package, subject to the same data-access restrictions.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Appendix A: Reproducibility Tables

Table A1 Temporal experiment settings

Parameter	Value
Raw observations	12,000
Sites	250
Assets	1,499
Primary target	Future Overheating OR Voltage Anomaly
Primary horizon	72 h
Sensitivity horizons	24, 48, 72, 96, 168 h
History	Current + up to 3 previous observations
Grouping	5-fold GroupKFold by Site_ID
Random seed	42
XGBoost estimators	250
XGBoost max depth	3
XGBoost learning rate	0.05
GRU hidden units	32
GRU epochs	25
Bootstrap replicates	3,000 (primary 72 h)

Table A2 72-h paired bootstrap and history sensitivity

Analysis	Value
Temporal − current pooled OOF ROC-AUC	0.07185
Site-level bootstrap mean difference	0.07413
95% CI	0.02346 to 0.12428
Bootstrap unit	Site_ID
Bootstrap replicates	3,000
≥1 prior observation gain	0.04481
≥2 prior observations gain	0.00001
≥3 prior observations gain	0.01437
≥3 prior observations bootstrap 95% CI (1,000 reps)	−0.06185 to 0.09188

Table A3 Additional robustness controls

Analysis	Value	95% CI / note
Mask-only minus current AUC	+0.0156	[−0.0085, 0.0395], 3,000 site bootstrap
Mask-only pooled ROC-AUC	0.5021	Current-state pooled AUC = 0.4868
7-day rolling minus current AUC	−0.0123	[−0.0517, 0.0239], 3,000 site bootstrap
7-day rolling pooled ROC-AUC	0.4745	Mean ± SD = 0.4764 ± 0.0570
Locked temporal pooled ROC-AUC	0.5587	Mean ± SD = 0.5592 ± 0.0295
Temporal minus mask pooled OOF gap	+0.0566	Descriptive; no new paired CI

Table A4 Asset-class event stratification at 72 h

Asset class	Eligible N	Future TE events	Future rate	Overheating events	Voltage anomalies
HVAC	2,049	20	0.98%	103 (4.21%)	8
Generator	1,775	36	2.03%	83 (3.86%)	10
UPS/Rectifier	1,802	33	1.83%	80 (3.75%)	13
DC Rectifier	1,650	27	1.64%	74 (3.67%)	10
RF Power Amp	1,508	23	1.53%	79 (4.40%)	0
Transfer Switch	1,207	17	1.41%	52 (3.57%)	6

Note: future-event counts use the 72-h eligible population (N = 9,991); overheating and voltage counts are recorded-event counts in the full 12,000-row source dataset. No class-specific ranking claim is made from these counts alone.

Table A5 Generalization stress-test results

Analysis	Value

Analysis	Value
Chronological split (Eq. 43)	70/30 by reference time, 72-h purge
Train / test N	6,530 / 2,999
Test events	55
ROC-AUC / PR-AUC / Brier	0.5015 / 0.0177 / 0.01829
LOCO ROC-AUC	DC Rectifier 0.4956; Generator 0.4972; HVAC 0.6042; RF Power Amp 0.6852; Transfer Switch 0.5660; UPS/Rectifier 0.5389
Gated variant (Eq. 19) vs temporal	ΔAUC −0.0392 [−0.0657, −0.0115]; Brier 0.01549 vs 0.01553