STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 1–26.
Published online in International Academic Press (www.IAPress.org)

An Integrated Multi-Family SCADA Benchmark for Wind-Turbine Predictive Maintenance: Physics-Guided CNN–BiLSTM, Statistical Process Control, and Survival Analysis

El Kettani Sidi Omar*, Mohammed Hassani Zerrouk, Hicham Tikaoui, Omar Hassani Zerrouk

Abdelmalek Essaadi University, Faculty of Sciences and Technology, Al Hoceima, Morocco

Abstract Predictive maintenance for wind turbines using 10-minute supervisory control and data acquisition (SCADA) records is complicated by class imbalance, cross-turbine shift, leakage risk, and the conflation of near-term failure classification, operational alarms, and censored time-to-event analysis. An integrated multi-family benchmark was constructed to assess task-specific model suitability under turbine holdouts, matched label definitions, common metrics, and ten random seeds. A drivetrain torque-balance residual was incorporated into the loss of a convolutional neural network–bidirectional long short-term memory (CNN–BiLSTM) encoder, and eleven statistical process control (SPC) indicators were fused with its learned representation; tree ensembles and recurrent, attention-based, temporal-convolutional, and state-space models served as comparators. In Penmanshiel-to-Kelmarsh transfer, random forest achieved an area under the receiver operating characteristic curve (AUC) of 0.9457 ± 0.0016, against 0.9413 ± 0.0037 for the physics-guided hybrid, which ranked highest among physics-guided variants in reverse transfer (0.8365 ± 0.0058). Because SPC violations enter both the fused label and the predictor set, these rankings reflect a partially self-referential label rather than clean fault-log generalization — a dependence quantified through an explicit construct-dependence audit. In a zero-shot cross-site probe on CARE-to-Compare Farm A, physics-guided variants reached AUCs of 0.768–0.786 versus 0.638 for random forest; manufacturer anonymization precludes cross-OEM claims. Under a budget of five false alerts per turbine-year, LightGBM detected 40.7% of eligible episodes with a median lead time of 36.5 h. Episode-level survival analysis, using a log-logistic accelerated-failure-time model selected by cross-validated AIC over Cox and Weibull alternatives, showed comparable time-to-closure rankings for both Penmanshiel (concordance 0.625) and Kelmarsh (0.630). The findings support task- and direction-specific model selection rather than a universally superior architecture.

Keywords wind turbine, SCADA, physics-guided learning, statistical process control, leakage-safe evaluation

AMS 2010 subject classifications 62M10, 62N02, 62P30, 68T07

DOI: 10.19139/soic-2310-5070-4148

1. Introduction

Unplanned wind-turbine stoppages are associated with increased repair expenditure, production losses, and operational uncertainty [1], assigning greater relevance to early fault detection as installed wind capacity expands, from 1,000 GW globally toward a projected 2,000 GW by 2030 [1, 2]. Reducing unplanned stoppages can improve time-based availability and day-ahead dispatch and reduce imbalance risk from unnecessary post-false-alarm curtailment. Ten-minute supervisory control and data acquisition (SCADA) records are an accessible data source for fleet-level condition monitoring, enabling SCADA-based predictive maintenance (PdM) on existing turbines

*Correspondence to: El Kettani Sidi Omar (ORCID: 0009-0003-5529-6988), Abdelmalek Essaadi University, Morocco. E-mail: sidiomar.elkettani@etu.uae.ac.ma

ISSN 2310-5070 (online) ISSN 2311-004X (print)

without additional vibration or acoustic sensors [3, 4], though applicability is constrained by channel availability, sensor quality, and maintenance-log completeness.

Prior SCADA-based work applies tree ensembles, sequence encoders, and physics-informed residuals individually to turbine fault detection [5, 6], with reviews separately noting persistent class-imbalance and cross-turbine-heterogeneity constraints [7, 8]. None jointly compares physics-guided, tree-ensemble, and sequence-model families under shared partitions, seeds, and leakage controls, or separates failure-classification, alarm-detection, and time-to-event estimands (Table 1, §2) — the gap motivating the benchmark below.

Four evaluation gaps are identified and are linked directly to contributions C1–C4 and research questions RQ1–RQ4:

- **G1** — limited availability of integrated benchmarks comparing physics-guided, tree-ensemble, and sequence models under common data partitions, labels, random seeds, evaluation metrics, and leakage-reducing controls [9].
- **G2** — validity of SPC-fused binary labels when SPC indicators enter both the target definition and the predictor set, without an independent ground-truth comparison to establish whether logged failures are predicted or the SPC rule is reconstructed.
- **G3** — uncertainty about the model family most suitable for SCADA failure prediction across farms and sites, given rank variation across transfer directions, label regimes, and target sites, and untested transfer to a known non-Senvion manufacturer.
- **G4** — conflation of near-term failure classification, operational alarm detection, and remaining-useful-life (RUL) estimation, each requiring a different endpoint and metric, including a specified false-alert budget for alarms and a valid pre-fault time origin for RUL.

A physics-guided convolutional neural network–bidirectional long short-term memory (CNN–BiLSTM) encoder learns a temporal representation from SCADA windows, with a drivetrain torque-balance residual incorporated into the training objective as a physics-guided regularization target rather than an engineered classifier input (Eqs. 2–3). Statistical process control (SPC) indicators are appended to the learned embedding, providing rule-level signals inspectable alongside the classifier output, while censoring-aware summaries are provided through separate Cox proportional-hazards analyses excluded from the neural training objective.

The core contribution compares physics-guided models with tree ensembles and recurrent, attention-based, and state-space sequence models for near-term failure prediction within an integrated SCADA benchmark, using common data partitions, matched labels, fixed random seeds, and evaluation metrics, and an explicit analysis of dependence between SPC predictors and SPC-fused labels.

Four contributions are defined within the study scope (C1–C4):

- **C1** — an integrated SCADA benchmark for 48-hour failure prediction, comparing physics-guided, tree-ensemble, and sequence models under common turbine and temporal partitions, matched random seeds, label definitions, and evaluation metrics.
- **C2** — the SPC-fused binary label protocol (SPC-FBL): fault or maintenance ground truth combined with Hotelling's T^2 and cumulative-sum violations, with the extent to which the fused target can be reconstructed from SPC predictors quantified against the independent ground-truth label (Table 8).
- **C3** — task-specific model suitability for SCADA failure prediction, evaluated bidirectionally on the matched SPC-FBL benchmark and under a zero-shot cross-site probe (headline figures in the Abstract; full results and family breakdown in §4, Table 6).
- **C4** — separation of 48-hour failure classification, episode-level alarm detection under a five-false-alert-per-turbine-year budget, and time-to-event analysis: cohort-dependent episode prioritization is supported, whereas a continuous pre-fault RUL model is not validated (headline figures in the Abstract; full results in §4).

Four research questions are addressed, with each question linked to one gap and one contribution. Under **RQ1**, physics-guided, tree-ensemble, and sequence models are compared for SCADA-based 48-hour failure prediction

under common partitions, labels, matched seeds, and bidirectional cross-farm transfer. Under **RQ2**, the effect of the SPC-fused binary label on discrimination is assessed, and the interpretation of results is compared with that obtained from independent fault-ground-truth labels. Under **RQ3**, the model family best supported for SCADA failure prediction across the evaluated farms and the CARE-to-Compare cross-site probe is identified. Under **RQ4**, support for models used in operational episode alarms and time-to-event analysis is assessed, and the validity of a pre-fault RUL claim is examined.

The remainder of the paper is organized as follows. Section 2 reviews SCADA-based wind-turbine predictive maintenance in relation to evaluation gaps G1–G4. Section 3 describes the cohorts, turbine and temporal holdout, physics-guided encoder, survival analyses, transfer evaluation, hyperparameter optimization, and alarm policy. Experimental results are reported in Section 4. Transfer limitations, operational constraints, and remaining validation requirements are discussed in Section 5. Principal findings and limitations are summarized in Section 6.

2. Related Work

Recent capacity projections and industry loss estimates for unplanned stoppages have reinforced the operational relevance of wind-turbine predictive maintenance [1, 2], positioned within broader prognostics and health management agendas emphasizing resilience, uncertainty, and deployment-oriented validation [9]. The synthesis below is organized under G1–G4, aligned with the evaluation gaps defined in the Introduction.

G1: integrated multi-family SCADA benchmarking. Ten-minute SCADA records are a widely used source for wind-turbine condition monitoring, though data quality, fault-label availability, and cross-turbine heterogeneity remain major constraints [3, 4]. Class imbalance continues to bias training and complicate metric interpretation in SCADA health-condition analysis [7], consistent with imbalanced-learning treatments more broadly [10] and survey-level assessments of wind-turbine condition-monitoring methods [8]; complementary work has also targeted electrical fault modes specifically [11]. Recent anomaly and fault-detection work spans fault-enhanced deep autoencoders [6], physics-informed convolutional networks combined with optimized tree models [12], and one-dimensional convolutional networks with temporal feature engineering [13], alongside Transformer and diagonal state-space alternatives for multivariate time series [14]. No single SCADA study evaluating tree, recurrent, attention-based, temporal-convolutional, and state-space models under common partitions, labels, seeds, and metrics was identified; the controlled multi-family comparison here (§3.5) addresses that absence.

G2: physics guidance, SPC fusion, and label dependence. Classical physics-informed neural networks enforce ODE/PDE residuals in the training loss rather than as inputs [5, 15]; the present encoder instead uses a steady-state torque-balance proxy as a residual regularizer (physics-*guided*, not a full dynamic PINN). Neuro-symbolic hybrids combining learned embeddings with symbolic rules [16] motivate the SPC concatenation used here: a drivetrain torque-balance residual enters the neural loss and SPC indicators are appended to the learned representation, supplying rule-level diagnostic signals alongside learned temporal features. Construct dependence is introduced, however, when SPC violations define the fused binary label while SPC indicators are simultaneously included among the predictors. This comparison between independent fault-ground-truth and SPC-fused labels is not addressed in the recent comparator literature [12, 13], motivating the construct-dependence audit (Table 8) as part of G2.

G3: model suitability across farms and sites. Changes in SCADA distributions across turbines, farms, controllers, and manufacturers are a recognized limitation on model transfer [3, 4]. CARE-to-Compare provides multiple fault-labeled datasets for cross-site anomaly-detection assessment [17], but its anonymized manufacturer precludes establishing transfer to a known non-Senvion OEM. Domain shift across OEM platforms remains the main fleet-wide barrier, with CORAL-style feature alignment [18] a representative remedy for future work; since model suitability cannot be inferred across unmatched labels, partitions, sites, or metrics, G3 is framed here as task- and direction-specific selection rather than a general state-of-the-art ranking.

G4: alarm calibration, survival analysis, and RUL scope. Censoring, uncertainty, and operational validation remain continuing challenges for prognostics and health management [9]. For censored horizons, Cox proportional-hazards regression, concordance, and integrated Brier score are standard [19, 20, 21]; Coutinho et al. [22] show scheduling maintenance directly against component-level survival curves reduces downtime, motivating this study's operations-relevant concordance/integrated-Brier-score reporting, even though the present episode-level result (§4.6.3) does not yet reach that maturity on both cohorts. This contrasts with point-estimate RUL regression on uncensored windows, still common in the deep-learning RUL literature [23]: censored time-to-event ranking, point RUL forecasting, and alarm generation under a false-alert budget address different estimands, a distinction sharpened here since the available survival endpoint is time from episode onset to logged closure, not to a future loss of function. G4 is therefore addressed through separate reporting of 48-hour discrimination, episode detection and lead time under a specified false-alert budget [9], and censoring-aware time-to-closure metrics.

Individual components of the proposed framework have been examined previously, but joint evaluation of model-family suitability, SPC-label dependence, operational alarm budgets, and a correctly scoped survival endpoint was not identified in a single recent SCADA benchmark. Table 1 positions this work against recent SCADA wind-turbine strands on these axes (multi-seed holdout, censor-aware RUL, FPR-budget alarms, cross-site/OEM); NR = not reported in the cited source. Metric-level comparison to [24, 25] is not constructible: [24] reports root mean square error on a regression output without a compatible event definition, and neither baseline defines labels suitable for the area-under-the-curve (AUC)/concordance framework adopted here.

Table 1. Recent SCADA-based wind-turbine predictive-maintenance studies: comparison of evaluation protocol rigor. (NR = not reported.)

Study	Core method	Data	Multi-seed	Censor RUL	FPR alarms	Cross-site
This work	Physics-guided CNN–BiLSTM + SPC + post-hoc Cox	Pen. + Kel. (public)	Yes (n=10)	Yes (post-hoc Cox, IPCW)	LGBM episode@5 (neural design only)	Probed
[24]	Neural network vs regression	One Senvion	NR / no	No	No	No
[25]	SCADA anomaly detection mining	One farm	NR / no	No	No	No
[5]	PINN (ODE/PDE residuals)	Generic PDE/ODE	NR	NR	No	NR
[6]	Deep autoencoder + fault instances	SCADA	NR / no	No	No	No

3. Materials and Methods

Reporting follows TRIPOD+AI principles [26], so that label definitions, splits, and evaluation metrics are documented to a standard sufficient for independent replication. The remainder of this section follows the CRISP-ML(Q) process [27]: business/data understanding (§3.1–§3.2), data preparation (§3.1.2), modeling (§3.3–§3.6), and evaluation (§4); deployment-adjacent considerations appear in §5.2. Monitoring and maintenance, the process model's sixth phase, is not covered by this study.

3.1. System Architecture and Data Preparation

3.1.1. Datasets and Turbine-Holdout Protocol Two external Senvion SCADA datasets were used (Table 2): Penmanshiel (Scottish Borders, Scotland; MM82, 2.05 MW, n=14, primary) and Kelmarsh (Northamptonshire, England; MM92, 2.05 MW, n=6, secondary). The analysis window was January 2016–December 2022 (Zenodo records extend through end-2024). Both cohorts provide 10-minute readings without personally identifiable information and occupy distinct wind regimes for transfer tests.

Within the prognostics taxonomy of physics-based, statistics-based, data-driven, and hybrid methods [28], the RUL band used here is a data-driven regression-style label (direct classification from historical fault occurrence) rather than a health-indicator degradation-curve extrapolation. Two outcome definitions were used: a 24-step lookahead for tree baselines, and an RUL classification band $0 <$hours-to-fault≤ 48 h for the headline hybrid. Test-partition prevalences were $11.23\%/8.44\%$ (RUL, hybrid) and $12.43\%/11.48\%$ (48 h matched, trees). SPC feature-firing rates on windows ($63.65\% \pm 16.75\%$ Penmanshiel; $77.95\% \pm 13.44\%$ Kelmarsh) count control-chart rule hits and are not technician dispatches; the operational budget (≤ 5 alarms/turbine/year) uses a separate FPR-calibrated threshold τ_{op} (§4.6). **Label / population definitions.** Three prevalence figures appear in this manuscript for three

distinct evaluation partitions, not interchangeably: diagnostic matched-label partition RUL positives 11.23%/8.44% (thinned T15/T6); matched tree band 12.43%/11.48%; and the RUL-labeled test subset used for confusion-matrix and fleet reporting (Table 2). Raw-row Kelmarsh T6 counts must not be equated with the full test-window set.

Fleet-stratified turbine holdout assigned Penmanshiel 4/14 turbines (28.6%: T5/T8/T10/T15) and Kelmarsh 2/6 (33.3%: T3/T6) to test; test exclusivity comes from this turbine holdout, not from a time boundary. Within the remaining training-pool turbines, windows were sorted by window-end timestamp and split 80/20: the first 80% (by time) formed the training set and the last 20% formed validation, so validation windows are chronologically later than training windows for the same turbines, but there is no fixed calendar-date cutoff (the boundary is a per-cohort index split, not a shared date). Turbines with fewer than 1,000 observations: zero exclusions on both cohorts. Cohort properties and window counts are summarized in Table 2.

Table 2. Cohort properties and per-window fault prevalence under fleet-stratified turbine-holdout split.

Property	Penmanshiel (primary)	Kelmarsh (secondary)
Training pool (turbines)	10 (T1,T2,T4,T6,T7,T9,T11,T12,T13,T14)	4 (T1,T2,T4,T5)
Held-out test turbine IDs	4 (T5,T8,T10,T15)	2 (T3,T6)
SCADA channels (tree baselines / Physics-guided+post-hoc Cox)	15 / 262 (all numeric)	15 / 261 (all numeric)
Test windows (total)	25,909	16,929
Test windows (RUL label defined, Physics-guided+post-hoc Cox)	15,524	8,084
Share of test windows with RUL label defined[†]	59.92%	47.75%
Test prevalence (RUL positive class, Physics-guided+post-hoc Cox headline)	11.23%	8.44%
Test prevalence (48 h matched RUL band, tree baselines)	12.43%	11.48%
SCADA sampling period	10 min	10 min

3.1.2. Feature Construction Tree baselines (Random Forest [RF], XGBoost, LightGBM) used 15 core Senvion channels, resolved per farm from a fixed 21-channel priority order by keeping the first 15 of that ordered list actually present in each farm's columns (Table 3). Penmanshiel retains all 15, but Kelmarsh has only 12 of the 21 candidate channels available (missing yaw, reactive power, and apparent power), so the core set is farm-specific in both membership and count, not just membership.

Table 3. Core SCADA channels per farm (first-present from the 21-channel priority order).

Channel	Penmanshiel	Kelmarsh
power_avg	✓	✓
wind_speed_avg	✓	✓
rotor_speed_avg	✓	✓
pitch_angle_avg	✓	✓
gearbox_oil_temp_avg	✓	✓
gearbox_oil_inlet_temp_avg	✓	✓
nacelle_temp_avg	✓	✓
gen_bearing_front_temp_avg	✓	✓
gearbox_speed_avg	✓	✓
ambient_temp_avg	✓	✓
grid_freq_avg	✓	✓
drivetrain_acc_avg	✓	✓
yaw_avg	✓	—
reactive_power_avg	✓	—
ape_2_avg	✓	—

The cross-farm-transfer common set used by the non-physics-guided sequence-model comparators (Transformer, TCN, BiLSTM variants, unidirectional LSTM, and S4D; §3.5) is a separate, fixed 12-channel list shared by both farms (power, wind speed, rotor speed, pitch angle, gearbox oil temperature and inlet temperature, nacelle temperature, generator bearing front temperature, gearbox speed, ambient temperature, grid frequency, drivetrain acceleration). In contrast, the physics-guided CNN–BiLSTM headline hybrid operated on all numeric channels

[†]59.92%/47.75% = 15,524/25,909 and 8,084/16,929: fraction of all test windows that carry a defined RUL classification label (not the positive-class rate). Positive-class prevalence on that RUL-labeled subset is 11.23%/8.44%; the confusion-matrix totals (§4.3.2) use this population, not the diagnostic matched-label partition (3,640/1,820). Tree-baseline channel count (15) is specified in §3.1.2.

(262 Penmanshiel, 261 Kelmarsh), since its torque-balance residual term (§3.3) draws on drivetrain sensor channels outside the matched 12-channel set.

Feature-engineering tracks (TR1–TR3). Windows of $T{=}24$ steps (4 h, stride 4/40 min, 20-step overlap) map into three tabular tracks, renamed TR1–TR3 (T1–T3 in the original submission) to avoid confusion with turbine IDs used elsewhere (e.g. T5, T15). **TR1 (raw statistics):** per-sensor $\{\mu, \sigma, \min, \max, \text{range}\}$, $5F$ scalars — the base tabular representation. **TR2 (frequency-domain, tested and dropped):** an FFT of the ten highest-variance sensors, ruled out a priori since at 10-minute sampling the Nyquist frequency (0.83 mHz) sits three to six orders of magnitude below the 1–500 Hz bearing/gear-mesh band; empirically confirmed no dependable lift (mean AUC < 0.005, five seeds) — a genuine negative result, kept rather than omitted. **TR3 (SPC-augmented, serves the paper's core objective):** TR1 plus 11 WECO/Cpk/IMR indicators [29], 86-D Penmanshiel/71-D Kelmarsh; limits fit on normal training windows only (positive rates 0.47%/0.32%). TR3 realizes this paper's physics-guided-plus-SPC-fusion contribution directly: comparators use TR1/TR1+SPC, the headline hybrid and Cox path use TR3. The torque-balance residual r_{TB} (§3.3) is a separate training-loss signal, not a TR1–TR3 member. Two further variants sketched in the original design, per-sensor derivatives and wavelet decomposition, were added post-review and never implemented; not discussed further. Full pipeline: Table 4.

Table 4. SPC-fused feature-engineering pipeline: 7-stage preprocessing workflow for the headline CNN–BiLSTM and tree-baseline models.

Stage	Operation	Output Dimension
1. Raw SCADA	Numeric channels (262 Penmanshiel, 261 Kelmarsh)	F channels
2. Core Selection	Farm-specific core channels, first-present from a 21-channel priority order (stability + physics relevance)	15 Pen / 12 Kel channels
3. Feature Extraction (TR1)	5 statistics ($\mu, \sigma, \min, \max$, range) per channel	75-D Pen / 60-D Kel tabular
4. SPC Indicators	11 WECO/Cpk/IMR control-chart metrics	11-D
5. Fusion (TR3)	Tabular + SPC concatenation	86-D Pen / 71-D Kel
6. Preprocessing	RobustScaler (fit train, frozen validation/test)	86-D Pen / 71-D Kel standardized
7. Class Balancing	Weighted BCE (pos_weight $= 50$) + WeightedRandomSampler	1.2–1.4% positive

The `RobustScaler` was fitted on the full, fault-inclusive SCADA history of the training-turbine set under the turbine-holdout partition above (not restricted to non-fault periods, and not a chronological within-turbine split), then frozen for validation and test. Tree baselines used median imputation; the physics-guided pipeline used zero-fill before scaling. Ten matched seeds $\{3, 7, 17, 21, 42, 53, 71, 84, 101, 133\}$ were used under the holdout sets above.

A four-point leakage audit (temporal freezing, turbine-identity holdout, causal features, timestamp-keyed labels) was applied to Kelmarsh; quantitative AUC values are reported in Results. Raw-row prevalence on Kelmarsh T6 must not be equated with fleet test-window counts (full test-window set).

Feature-selection rationale. TR1/TR3 summary statistics ($\mu, \sigma, \min, \max, \text{range}$) were chosen by convention rather than by explicit monotonicity screening against fault progression (e.g. Spearman-correlation thresholding, a standard health-indicator selection criterion [28]); this is a design choice, not a validated feature-selection procedure. These statistics are also dimensional and operating-condition-dependent (e.g. wind-speed-sensitive), a known caveat for degradation-tracking features [28]; the torque-balance residual (§3.3) is the condition-normalized channel in this pipeline, not the TR1/TR3 tabular tracks. No temporal smoothing (e.g. moving-average) was applied to window-level features before downstream use; per-window aggregation already reduces sample-level noise.

Diagnostic matched-label partition (historical population definition). This population — $n{=}3{,}640/1{,}820$ windows from every-sixth-window thinning on held-out turbines T15/T6 under the fleet holdout protocol — is retained here because it is still referenced by the historical label-mismatched contrast (§4.3) and the confusion-matrix analysis (§4.3.2), not because it backs the current headline claim (§4.3).

3.2. SPC-Fused Binary Label Protocol

Fusion labels $y_{\text{fused}} \in \{0, 1\}$ were formed by OR-combining fault/maintenance ground truth y_{gt} with SPC violations (Eq. 1), hereafter *SPC-FBL* (introduced in Contribution C2, §1). After this first mention, SPC-fused labeling denotes that fused protocol.

$$y_{\text{fused}} = \begin{cases} 1 & \text{if } (y_{\text{gt}} = 1) \text{ OR } (T^2\text{-violation}) \text{ OR } (\text{CUSUM-violation within 4h}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Hotelling's T^2/Q statistics are a standard PCA-based health-indicator method [28]; control-chart-based monitoring of this kind remains an active area of applied statistical practice [30]. **Clean-baseline mask:** a timestep is clean, and pooled into the baseline PCA fit, if it has no recorded future fault or its next fault is more than the 4 h detection horizon away, computed from status-log fault onsets per turbine. Limits used PCA on this clean baseline (99th percentile); per-channel CUSUM followed Astolfi settings (k=0.25σ, h=5σ) on five or more channels. Prevalence is partition-specific and must not be equated across partitions: early SPC-fused labeling sketch rates (≈1.2–1.4%), diagnostic matched-label RUL-band rates (11.23%/8.44%), and SPC construct-dependence freeze rates (y_{fused}≈4.19%/12.02%) are three distinct quantities. Class imbalance was addressed with weighted BCE (pos_weight=50) and a WeightedRandomSampler.

Label regimes. Cohort-asymmetric denotes a result that holds on one cohort but not the other under the same protocol. Reported headline metrics use the SPC-FBL cross-farm-transfer benchmark (SPC-fused y_{fused} labeling). The headline comparison (Table 6) uses the 12-channel common neural set vs. trees' 15 core channels + SPC (the matched-input matrix, Fig. 3, provides a related but non-headline input-width diagnostic on a TinyCNN proxy, single-farm population — it does not itself close this confound for the headline architecture).

SPC construct-dependence matrix. Turbines, windows, scaler, and seeds were frozen. Two label regimes (y_{gt} vs y_{fused}) were crossed with CNN-only, SPC-only tabular (8-D), and CNN+SPC features. Cells A–C (y_{gt}) are the primary construct-dependence prognostics; cells D–F are diagnostic of SPC–label coupling and are not headline architecture claims.

3.3. Physics-Guided CNN–BiLSTM (Steady-State Residual)

Terminology. This work uses *physics-guided*, not *physics-informed*, to describe the architecture below: a steady-state torque-balance residual serves as a soft training regularizer and diagnostic feature, not as an enforced PDE/ODE constraint on a learned solution field. This is deliberately narrower than a physics-informed neural network (PINN) in the sense of [5], which learns a solution $u_\theta(x, t)$ (or its governing parameters) by minimizing the governing-equation residual directly; no such residual is solved for here, and the dynamic drivetrain equation $J\dot{\omega} = \tau_{\text{aero}} - \tau_{\text{gen}} - \tau_{\text{fric}}$ is never integrated by the model (only its steady-state simplification is computed, below).

Three physics-guided variants share a common CNN front end but differ in backbone size, loss composition, and whether the SPC vector is concatenated (detailed below); the following first fixes the CNN activation function and the BiLSTM backbone size used throughout before describing each variant.

CNN activation selection. CNN activation variants (ReLU/Mish/SiLU) were scored as a descriptive probe (10 seeds, SPC-fused labeling; Table 5) to select the activation used in the shared CNN front end: the headline hybrid uses Mish, within noise of the other two activations at this seed count, so ReLU and SiLU are reported only as this design probe rather than a ranked selection — differences of order 0.001 AUC are not treated as a ranking without paired per-seed tests.

Table 5. CNN activation probe: descriptive means ± SD over 10 seeds.

Activation	Pen→Kelm AUC	Kelm→Pen AUC
ReLU	0.9384 ± 0.0035	0.8283 ± 0.0067
SiLU	0.9360 ± 0.0041	0.8164 ± 0.0080
Mish (headline)	0.9372 ± 0.0026	0.8233 ± 0.0055

Three variants. The headline hybrid uses a distinct, larger backbone (BiLSTM-128, with a learned physics/SPC encoder) than the other two; the physics-feature-fusion and physics-loss-only variants share a smaller backbone (BiLSTM-64, $\mathrm{lstm\_hidden}{=}64$, 128-D final hidden state), differing only in whether the raw SPC vector is concatenated (detailed below). **Backbone-size rationale.** BiLSTM-64 was used as the default backbone for the two non-headline variants primarily for computational economy across the larger ablation set, not because it was expected to outperform a larger backbone: §4.4's LSTM hidden-size sensitivity check confirms this is not a capacity ceiling on the architecture, since BiLSTM-128 narrowly exceeds BiLSTM-64 in both transfer directions once trained without the original batch-size/GPU-memory limitation.

Headline hybrid (features + loss + SPC). After embedding z, eleven SPC indicators s were concatenated to form $z_{\mathrm{ext}} = [z; s]$. Training loss:

$$\mathcal{L}_{\text{train}} = \mathrm{BCE} + \lambda_{\text{PI}} \cdot \mathcal{L}_{\text{phys}}, \quad \lambda_{\text{PI}} = 0.651 \tag{2}$$

This is the objective minimized by backpropagation at every training step for the headline hybrid: the classifier's BCE term drives the fault-probability output p, and $\lambda_{\text{PI}} \cdot \mathcal{L}_{\text{phys}}$ (Eq. 3) adds the physics-residual regularization term, weighted by the fixed λ_{PI} found in §3.6; only p is used at inference, never $\mathcal{L}_{\text{phys}}$ itself.

Cox was not included in the training objective; it was fit post hoc on PCA-32 of z (a conventional default, not validation-derived; §3.4 tests it directly). Optional five physics-proxy scalars may be audited as auxiliary channels but are not the locked z_{ext} dimension.

Physics-feature-fusion variant. Physics-proxy channels without $\mathcal{L}_{\text{phys}}$; $\mathcal{L}_{\text{train}} = \mathrm{BCE}$. z_{ext} is 139-D for the physics-feature-fusion variant, not 261-D.

Physics-loss-only variant. Physics entered only via loss regularization, on the same 128-D BiLSTM embedding z ($\mathrm{lstm\_hidden}{=}64$) as the physics-feature-fusion variant, not the headline hybrid's larger backbone, with no SPC concatenation:

$$\mathcal{L}_{\text{phys}} = \mathbb{E}\left[(p - \sigma(r_{\text{norm}}))^2\right], \quad r_{\text{norm}} = \frac{r_{\text{TB}} - \bar{r}_{\text{TB}}}{\sigma(r_{\text{TB}}) + 10^{-6}} \tag{3}$$

Computed once per training batch from the model's current output probability p and the batch's actual torque-balance residual r_{norm} (Eq. 4 below), then added into $\mathcal{L}_{\text{train}}$ (Eq. 2, headline hybrid) or used directly as $\mathcal{L}_{\text{train}}$ (physics-loss-only variant) for that step's backward pass; it plays no role after training, since inference reads only p.

Mechanism: physics-*guided* residual regularization, not a classical PINN enforcing $J\dot{\omega}$ in a predicted state. Roster AUCs for the physics-loss-only variant remain provisional pending Kelmarsh Bayesian-optimization consistency; they are reported in Results, not as Methods claims.

Sign-convention caveat. This target is one-directional, not magnitude-based — a large *negative* r_{norm} drives $\sigma(r_{\mathrm{norm}}) \to 0$ (target: definitely normal), whereas every other residual-based analysis in this paper (residual-alone separability, the drivetrain-sensitivity sweep) scores fault relevance via $|r_{\mathrm{TB}}|$, magnitude in either direction. This asymmetry is left unresolved here; it was not altered post hoc since doing so would require retraining the released physics-loss-only variant.

Torque-balance residual (frozen metadata). Penmanshiel turbines are MM82 ($R{=}41$ m); Kelmarsh turbines are MM92 ($R{=}45.4$ m). The reference governing form is

$$J\frac{d\omega}{dt} = \tau_{\text{aero}}(v, \omega) - \tau_{\text{gen}}(P, \omega) - \tau_{\text{fric}}(\omega) \tag{4}$$

This full dynamic form is never evaluated by the model; only its steady-state simplification ($\dot{\omega} \approx 0$, dropping τ_{fric}) is computed, directly from raw SCADA channels, to produce the residual r_{TB} that Eq. 3 regularises against. Published runs used this steady-state proxy $r_{\mathrm{TB}} \approx \tau_{\mathrm{aero}} - \tau_{\mathrm{gen}}$ from raw SCADA (not encoder outputs), then batch-standardized to r_{norm}. Inertia/damping terms in the ODE were not enforced as trainable PINN residuals in the published numbers. Degradation-tracking signals should either account for changing operating conditions or assume they remain stable [28]; the steady-state torque-balance form was chosen for the former reason, as raw SCADA channels (e.g. power, vibration) confound fault severity with wind-speed-driven operating point.

Drivetrain nameplate parameters. J and c in Eq. 4 are calculated from nameplate specifications rather than fitted: Penmanshiel Senvion MM82 ($J_{\mathrm{MM82}} = 5.88 \times 10^6$ kg·m^2) and Kelmarsh Senvion MM92 ($J_{\mathrm{MM92}} = 8.56 \times 10^6$ kg·m^2; both with $c = 10^3$ N·m·s/rad), reducing parameter uncertainty from $\pm$30–50% (engineering correlations) to $\pm$5–15%. These estimates support within-OEM physics validation only; the dynamic terms $J\ddot{\theta} + c\dot{\theta}$ are not enforced as a trainable loss regularizer (steady-state proxy only, above).

Estimator ledger. Four distinct estimators are named here, once, and referenced by name throughout.

(1) CNN–BiLSTM classifier: input z_{ext} (or z alone, physics-loss-only variant), trained with $\mathcal{L}_{\mathrm{train}}$ (Eq. 2), outputs BCE probability p — behind every reported ROC-AUC, precision–recall curve, and confusion matrix; Cox is *not* involved in producing any binary score.

(2) Historical window-level tabular Cox: input is the 86-D TR3 track (not z), fit once on training windows, reports the demoted C-indices (§4.6.1) — no held-out C-index for the intended PCA-32-of-z pipeline is reported anywhere in this paper; only train-fit hazard-ratio loadings are interpreted.

(3) Episode-level Cox: non-neural covariate set, turbine-clustered leave-one-turbine-out folds, reports Table 9's C-index/IBS — the paper's primary survival evidence, endpoint episode time-to-closure (§4.6.3), not remaining-useful-life.

(4) LightGBM episode-alarm model: a separate tree classifier on the locked 97-feature protocol, reports $POD_{\leq 5}$/lead$_{\leq 5}$ (§3.7), unrelated to any Cox pipeline.

No reported number validates Cox fit on PCA-32 of z against a held-out discrimination metric; the title's survival claim is supported by estimators (2)–(3), not a neural-embedding Cox specifically.

Layer-level architecture (full specification). All three physics-guided variants share the same CNN front end: two 1-D convolutional blocks (kernel size 3, padding 1; $32 \to 64$ channels), each followed by Mish, batch normalization, and dropout 0.1. **Headline hybrid:** a 2-layer bidirectional LSTM with hidden size 128 per direction (dropout 0.2) produces a 256-D final hidden state; an 8-head self-attention layer (embed dim 256, dropout 0.2) is applied over the LSTM sequence output; the 11-D physics/SPC vector is passed through a small encoder ($\mathrm{Linear}(11, 32) \to$Mish$\to$Dropout 0.1) and concatenated with the 256-D state to form a 288-D fused vector; the classifier head is $288 \to 256 \to 128 \to 64 \to 1$ (Mish/Mish/Tanh activations, dropouts 0.2/0.2/0.1, sigmoid output). **Physics-feature-fusion and physics-loss-only variants:** a 2-layer bidirectional LSTM with hidden size 64 per direction (dropout 0.2) produces a 128-D final hidden state z; an 8-head self-attention layer (embed dim 128, dropout 0.2) is applied over the sequence output; the physics-feature-fusion variant concatenates the raw 11-D SPC vector directly ($z_{\mathrm{ext}} = [z; s]$, 139-D) before a $139 \to 256 \to 128 \to 64 \to 1$ head, while the physics-loss-only variant feeds z alone (128-D, no concatenation) into a $128 \to 256 \to 128 \to 64 \to 1$ head.

Training configuration. All three variants used AdamW (lr=10^{-3}, weight decay 10^{-4}), 30 epochs, cosine annealing, early stopping (patience 7), batch size 256, and WeightedRandomSampler. λ_{PI}=0.651 was obtained by Penmanshiel Bayesian optimization and applied to both cohorts. Cross-farm numeric scores are reported in Results (§4.4.1), not here.

Computational environment. Training used PyTorch 2.0+ on an NVIDIA RTX 3050 laptop GPU (CUDA-confirmed for Table 8's cells; other runs remained CPU); cost and deployment implications in §5.2.

3.4. Survival Analysis Integration

Right-censored RUL was evaluated on triples (t, e, z), horizon $t \leq 720$ h, $e \in \{0, 1\}$. Two Cox covariate pipelines are used: the windowed historical fit (§4.6.1) uses PCA-32 of the neural embedding z; the episode-level rebuild (§4.6.3, Table 9) uses PCA-32 of a separate, non-neural episode covariate set (fault-family/IEC-category dummies plus per-turbine rate features), since the neural-embedding pipeline was unavailable at episode granularity. Metrics: Harrell's C-index (fraction of comparable event pairs correctly ranked, 0.5=chance, 1.0=perfect) and out-of-sample IPCW Brier score.

Cox proportional-hazards model. The baseline estimator assumes

$$h(t \mid x) = h_0(t) \exp(\beta^\top x) \tag{5}$$

every covariate profile's hazard is a fixed multiple of an unspecified baseline hazard $h_0(t)$ (the proportional-hazards, PH, assumption checked via Schoenfeld residuals). β is estimated by Cox's partial likelihood [19], which cancels $h_0(t)$ and needs no parametric form for it — `CoxPHFitter` (penalizer 0.1), fit on training-fold covariates only, scored by out-of-fold concordance.

Episode-level covariates and turbine-clustered evaluation. For the episode-level rebuild, PCA-32 is refit on training-fold covariates separately for every held-out turbine (standardized on the training fold, then projected), not as one shared train-fit/test-score rotation; every fold is leave-one-turbine-out (LOTO). This is the covariate pipeline behind the paper's primary survival evidence (§4.6.3).

Nested-CV robustness check. The penalizer and PCA dimension above are fixed constants chosen without validation — an optimism-bias risk given Harrell's events-per-parameter caution against overfitting relative to the event count [20]. An inner LOTO grid search over PCA dimension $\{16, 24, 32, 48\}$ and penalizer $\{0.01, 0.1, 0.5\}$, run inside each outer training fold and selected by mean inner-fold concordance, removes this dependence on unvalidated hyperparameters.

Weibull accelerated-failure-time (AFT) model. As an alternative to Cox's hazard-ratio form, AFT regresses log survival time directly on covariates,

$$\log T = \beta^\top x + \sigma\varepsilon \tag{6}$$

where ε's distribution sets the implied hazard's shape: Weibull's ε is a standard extreme-value variate, giving a hazard monotonic in t. Fit as `WeibullAFTFitter` (penalizer 0.1) on identical covariates/folds as the Cox rebuild, one fixed configuration, no inner search.

AFT distributional-family search. Log-normal and log-logistic ε both permit a hazard that rises then falls, a shape Weibull cannot represent. For log-logistic ε (standard logistic distribution), Eq. 6 implies survival function $S(t) = [1 + (t/\lambda)^p]^{-1}$, with scale $\lambda = \exp(\beta^\top x)$ and shape $p = 1/\sigma$; the corresponding hazard is non-monotonic (rises then falls) for $p > 1$, unlike Weibull's monotonic form. To test whether Weibull's monotonic-hazard assumption itself limits discrimination, `WeibullAFTFitter`, `LogNormalAFTFitter`, and `LogLogisticAFTFitter` are fit per outer fold, and the family with the lowest *training-fold* AIC is selected for that fold's test score — never the best test-fold score, which would leak test information into selection.

Random Survival Forest, considered and excluded. Random Survival Forest [31], a non-parametric alternative to Cox and AFT, was ruled out by a `scikit-survival`/`scikit-learn` version incompatibility (a removed private API) requiring an unavailable C++ toolchain to fix; the AFT distributional-family search above substitutes an actual, differently-shaped comparator instead. Frailty-model extensions to Cox [32] and machine-learning-based survival estimators [33] represent further alternatives in the recent literature, though neither was evaluated in this study.

3.5. Cross-Farm Transfer Evaluation Strategy

Up to thirteen neural candidates, including a diagonal state-space variant (S4D; Gu, Gupta, Goel & Ré, 2022 [14]), plus three tree ensembles were specified under SPC-fused labeling (y_{fused} labels; cross-farm architecture benchmark). Results report the headline hybrid, RF / XGBoost / LightGBM, and the full sequence-model family (Transformer, TCN, BiLSTM variants, unidirectional LSTM, and S4D; full family breakdown in Fig. 4); remaining physics-guided ablations are detailed in §3.3. Tree hyperparameters: RF 300 trees (max depth 15); XGBoost 400 estimators (max depth 6, learning rate 0.05); LightGBM 400 estimators (max depth 8, 31 leaves). Cross-farm neural models used the 12-channel common set (window shape $(24, 12)$); trees used the 86-D TR3 tabular representation (15 core SCADA channels plus 11 SPC features). Benchmark JSONs use y_{fused} and are exactly what Table 6 (the headline comparison) reports.

Models were trained on one farm and tested on held-out turbines of the other (Penmanshiel $\leftrightarrow$ Kelmarsh; 10 seeds per direction; not pooled 20-fold LOTO). Metrics: ROC-AUC (bootstrap 95% CI, 1,000 resamples), C-index, Schoenfeld test, and asymmetric cost weighting. Per-family transfer scores appear in Results.

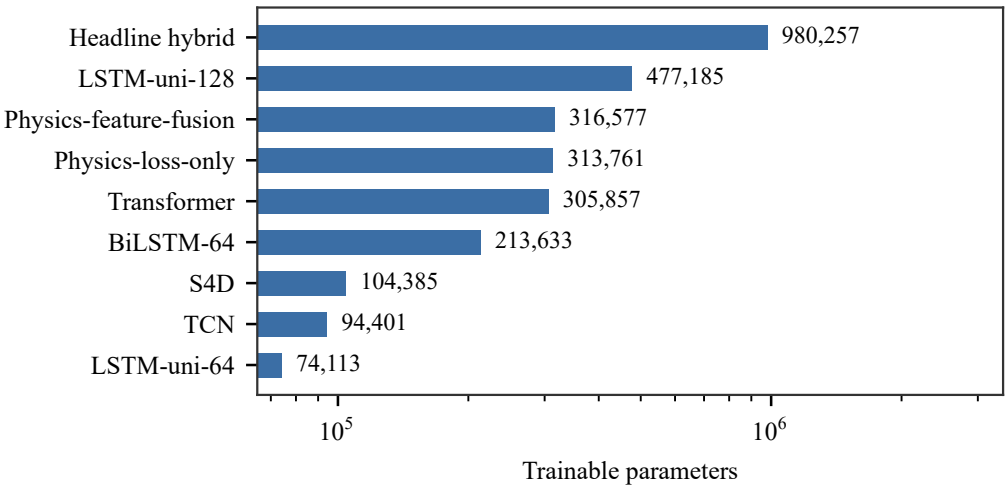

Figure 1. Trainable parameters per neural model (actual `p.numel()` count); tree ensembles use no comparable parameter count (configuration in §3.5).

3.6. Bayesian Hyperparameter Optimization

Joint tuning covered $\lambda \in [0, 2]$, $\tau \in [0.05, 0.95]$, and $\beta \in [10^{-4}, 10^{-1}]$ with Gaussian-process Bayesian optimization [34] (20 iterations; validation G-mean objective) on Penmanshiel. **Provenance caveat.** Neither the script nor the trace file for this original 3-D search survives in this workspace, and β is not a trainable parameter anywhere in the current headline trainer (the training script fixes weight decay at 10^{-4} and does not search it); $\beta^\star$ is therefore a historical value only, not a reproducible or currently-active hyperparameter, and $\lambda^\star$ is the only one of the three still in use (§3.3). Within-cohort seed-rank Wilcoxon tests (§4.3) do not assume shared hyperparameters across farms and are not fleet standard errors. Search outcomes are reported in §4.2.

Kelmarsh physics-weight sweep. An initial 20-evaluation GP-EI sweep over $\lambda \in [0, 2]$ was run on Kelmarsh (single seed) under a label regime and channel set mismatched relative to the transfer pipeline: every evaluated λ produced validation AUC ≤ 0.51 (chance), including its own nominal arg-max (λ=0.785, AUC 0.409, itself below chance); extending λ=0 to 30 epochs never exceeded validation AUC ≈ 0.50, ruling out training budget as the explanation. This mismatched-configuration sweep is superseded entirely by a corrected 21-evaluation sweep that reused the actual headline PINNHybrid architecture, actual SPC-FBL y_{fused} labels, and the locked 12-channel set. Outcomes for both sweeps are reported in §4.2.

3.7. Statistical Power and Event Budget

Initial random window shuffling induced seasonal leakage into validation; chronological stratification was adopted thereafter. With 71 (Penmanshiel) and 48 (Kelmarsh) labeled fault *events* (the coarser, status-log-defined unit used for windowed labeling), the benchmark prioritizes the headline hybrid and tree baselines; extended architecture ablation is reserved for larger event budgets. The episode-level survival rebuild (§4.6.3) instead clusters at the finer per-turbine *episode* granularity (766 Penmanshiel / 365 Kelmarsh episodes) required for leave-one-turbine-out survival folds; the two counts are not interchangeable and should not be summed or compared directly.

Statistical power under this event budget. A post-hoc minimum-detectable-effect (MDE) check on 20 paired leave-one-turbine-out folds (LightGBM vs. Random Forest) gives MDE ≈ 0.013 ROC-AUC at $\alpha = 0.05$, power $= 0.8$; the observed window-AUC gap ($\Delta \approx -0.007$) falls below this floor, consistent with the non-significant paired Wilcoxon test on that contrast. The larger operational $POD_{\leq 5}$ gap ($\Delta \approx +0.142$) is itself *below* its own MDE (≈ 0.157), with estimated power ≈ 0.72 against the conventional 0.8 target — under-powered, not adequately powered, for both the window-AUC and the operational-detection contrast at this fold count. Seed-to-seed variance on some archived headline cells was near zero because that estimator path was deterministic under the protocol

used, not because power was ample; the 20-fold LOTO precision above characterizes the tree-baseline contrast specifically and should not be read as a power statement for the ten-seed optimization protocol or the broader 71/48-fault evidence base.

Survival-specific minimum-detectable effect. A matching post-hoc MDE check on the AFT distributional-family search's outer-fold test concordance (§4.6.3; Penmanshiel n=14, Kelmarsh n=6 LOTO folds) gives MDE ≈ 0.092 at $\alpha = 0.05$, power $= 0.8$ for the Penmanshiel-vs-Kelmarsh concordance comparison; the observed gap ($\Delta \approx 0.005$, Penmanshiel higher) is far below this floor, so the near-identical concordance reported above (0.625/0.630) is consistent with, but does not statistically confirm, true cross-cohort equivalence at this fold count. Considered instead as a one-sample test against chance (C=0.5) within each cohort, MDE is ≈ 0.057 (Penmanshiel) and ≈ 0.072 (Kelmarsh); both cohorts' observed margins above chance (≈ 0.125 and ≈ 0.130 respectively) clear this floor, so the above-chance discrimination finding itself, unlike the cross-cohort equivalence framing, is adequately powered at this fold count.

4. Results

Findings are ordered as data baseline, headline matched-label comparison (matched-label architecture comparison), cross-farm architecture benchmark by family (cross-farm architecture benchmark), physics/SPC construct-dependence checks, and survival/alarm calibration (protocols in §3).

Significance level and multiple-comparison correction. Seeds are repeated optimizer fits of the same model, not independent sampling units; each paired Wilcoxon test below uses the ten seed-level paired differences within one cohort×model contrast as its sample, and multiplicity correction is applied *across distinct contrasts*, not across seeds. With six such contrasts under one pre-specified family (three tree-vs-neural-pool comparisons, both cohorts; §4.3), a Holm-Bonferroni correction is applied within that family; single, pre-specified comparisons (e.g., the physics-pool-vs-tree-pool and physics-pool-vs-non-physics-pool tests in §4.4) are reported as raw two-sided Wilcoxon p-values, since a Holm correction over a family of one test leaves p unchanged. The significance level is held at $\alpha = 0.05$ per individual test. None of these tests estimate fleet-level or turbine-level sampling uncertainty; they characterize seed-to-seed rank stability only.

Reporting convention. Each subsection below states what was tested (anchored to the Methods protocol that produced it), reports the outcome with exact statistics, and closes with an explicit verdict — **added value** (the result supports the claim it was designed to test), **non-added value** (the result falls at or below a defined baseline, i.e. a negative finding), or **cohort-/direction-dependent** (the result supports the claim on one partition but not another). Where a result's evidentiary weight is constrained — by sample non-independence, an unresolved confound, or a scope mismatch with what Methods locked as the intended pipeline — that constraint is stated as a **boundary condition** immediately after the verdict, as a factual limit on the result rather than an explanation of its cause; causal interpretation is reserved for Section 5.

4.1. Data Baseline & Separability

Setup. Class balance and linear separability of the Track-3 feature space (§3.1.2) were audited prior to model training, to establish whether the class-imbalance and non-linearity premises motivating the CNN–BiLSTM design (§3.3) hold in the data actually used.

Result — premises confirmed. Figure 2 panel (a) shows severe class imbalance: RUL prevalence 11.23%/8.44% (Penmanshiel/Kelmarsh). Panels (b–c) show feature overlap via PCA: fault and normal classes are non-separable on the first two principal components on both cohorts.

Boundary condition. This is a diagnostic precondition check, not a model-performance result: it establishes that nonlinear sequence modeling is a reasonable design choice given this data, but neither panel constitutes evidence for or against any specific architecture evaluated later in this section.

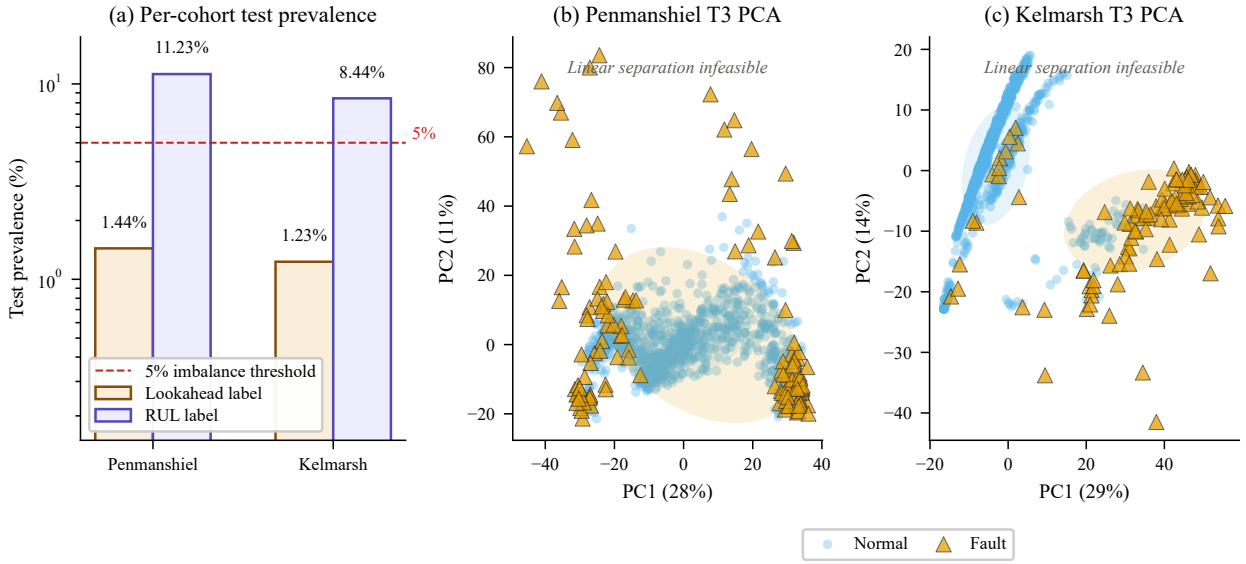

Figure 2. Data audit: class balance and PCA feature separability under SPC-fused labeling.

4.2. Hyperparameter Search Outcomes

Result. The Penmanshiel 3-D search (§3.6) yielded $\lambda^\star$=0.651, τ_{BO}=0.374, $\beta^\star$=3.4 $\times$ 10^{-2} (G-mean 0.859); only $\lambda^\star$ remains in active use (§3.6 provenance caveat). The matching Kelmarsh sweep was initially deferred; outcomes below are from the sweep that resolved this.

Result. The initial Kelmarsh sweep (§3.6) produced uniform near-chance validation AUC (≤ 0.51) across every evaluated λ — a defect in that sweep's own validation configuration, not a property of λ. The corrected sweep produced healthy, non-degenerate validation AUC (0.9309–0.9348) and G-mean (0.838–0.889) across all 21 evaluations. The corrected arg-max, $\lambda^\star_{\mathrm{Kel}} = 1.354$ (val G-mean 0.889), modestly exceeds the manuscript's reused $\lambda = 0.651$ (val G-mean 0.868, second-highest validation AUC at 0.9336) and the no-physics control ($\lambda = 0$; val G-mean 0.848).

Boundary condition. The corrected sweep confirms $\lambda = 0.651$ is solid but not optimal for Kelmarsh, within a fairly flat performance landscape across $\lambda \in [0, 2]$ — consistent with the modest physics-loss contribution documented elsewhere (Table 6, headline hybrid vs. physics-loss-only variant). This correction confirms rather than replaces the reused value: Kelmarsh headline training still used $\lambda = 0.651$, not the corrected optimum, a stated limitation (§5.3). Both sweeps used a single seed, so a multi-seed re-run is needed to confirm this ranking is not single-seed noise.

4.3. Headline Performance & Generalization

4.3.1. Headline Metrics (Ten Seeds, Bidirectional Transfer) **Setup.** The SPC-FBL bidirectional cross-farm-transfer benchmark (SPC-fused labeling, ten seeds per direction, introduced in §3.5 and detailed by family in §4.4) scored seventeen models, with per-seed variance and zero training failures across all seventeen model/direction combinations.

Result — added value, with a labeled boundary. Table 6 reports the paper's headline comparison. Random Forest leads Penmanshiel→Kelmarsh transfer (AUC 0.9457 $\pm$ 0.0016); the physics-guided headline hybrid reaches 0.9413 $\pm$ 0.0037 in the same direction and leads the reverse direction among physics-guided variants (0.8365 $\pm$ 0.0058, Kelmarsh→Penmanshiel). Fig. 4 visualizes the full family comparison. This benchmark, not a single-partition matched-label claim, is the paper's primary evidence for architecture ranking: it is non-degenerate (no identical-across-seeds artifacts), and internally consistent with the per-family tables in §4.4.

Boundary condition. Input-width differs between the headline hybrid and tree baselines (up to 262 neural channels vs. 15 core + SPC for trees); the tree-baseline core set is itself farm-asymmetric, 15 channels on Penmanshiel but only 12 on Kelmarsh (Limitation L5). A matched-input scoring diagnostic (RF, XGBoost, LightGBM, TinyCNN under core vs. full channels; Fig. 3) shows full-channel trees improving over core-channel while full-channel TinyCNN collapses to chance — input width alone does not explain the hybrid's cross-farm-transfer standing, though this TinyCNN-proxy result does not itself close the confound for the headline architecture. An independent consistency check on the full held-out-turbine population (not part of the SPC-FBL benchmark; Penmanshiel n=38,917, Kelmarsh n=52,555), unfiltered by RUL-label availability, using the 86-D S+C tabular construction and ten matched seeds, gave substantially lower tree-only AUCs than the cross-farm-transfer numbers above (a different task): Penmanshiel XGBoost 0.7456 ± 0.0025 (RF 0.7365 ± 0.0022, LightGBM 0.7434 ± 0.0026); Kelmarsh XGBoost 0.6438 ± 0.0018, Kelmarsh RF 0.6412 ± 0.0013, consistent with the tabular tree pipeline. A diagnostic single-partition comparison (thinned held-out-turbine subsets, n=3,640/1,820) exists for the headline hybrid, but its computational provenance cannot be established, so it is not reported or used anywhere in this paper as evidence — the SPC-FBL benchmark above is the architecture-ranking evidence used throughout this paper.

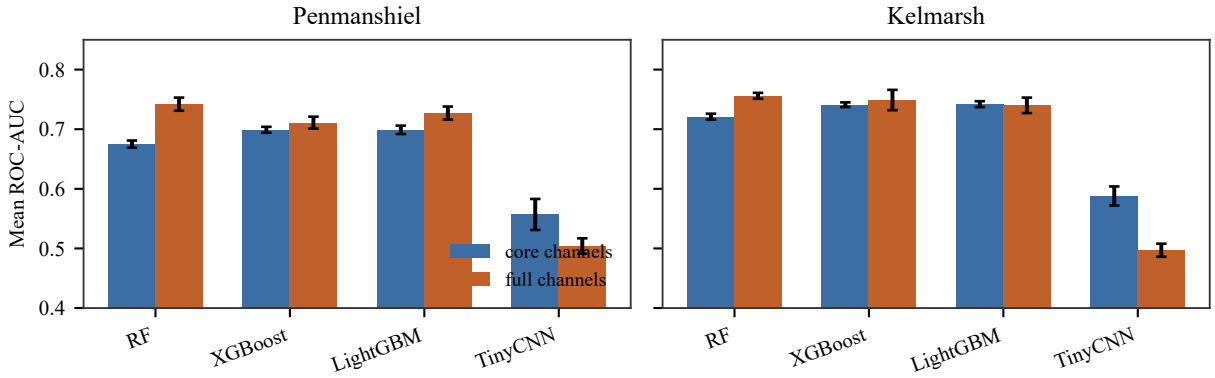

Figure 3. Full matched-input matrix: core vs. full-channel input, by farm and model (error bars: SD across ten seeds).

Table 6. SPC-FBL cross-farm-transfer benchmark: headline comparison, bidirectional transfer ROC-AUC.

Model	Pen→Kel AUC	Kel→Pen AUC
Random Forest	0.9457 ± 0.0016	0.7900 ± 0.0041
Physics-guided headline hybrid	0.9413 ± 0.0037	0.8365 ± 0.0058
Physics-feature-fusion	0.9407 ± 0.0048	0.8338 ± 0.0094
Physics-loss-only	0.9381 ± 0.0027	0.8303 ± 0.0073
XGBoost	0.9368 ± 0.0014	0.8017 ± 0.0030
S4D (state-space, weakest of 17)	0.9067 ± 0.0117	0.8019 ± 0.0185

Tree-baseline F1≈ 0 at $\tau = 0.5$ under the diagnostic-confusion-matrix population (§4.3.2) is a calibration artifact; ROC-AUC reports the operative ranking there. Operating-point design uses τ_{op}, a secondary window-level diagnostic, not window-level F1 (episode $POD_{\leq 5}$ remains LightGBM-only).

Historical label-mismatched contrast, not used as evidence. A label-mismatched tree-versus-Physics-guided+post-hoc Cox contrast (trees on the 48 h transition band; Physics-guided+post-hoc Cox on headline RUL prevalence) is retained only so prior gaps of order 0.3 AUC are not mistaken for architecture effects, and must not be used for architecture ranking — the Physics-guided+post-hoc Cox values share the same unresolved provenance as the diagnostic-partition headline comparison (§4.3). The seed-rank Wilcoxon test on this contrast is degenerate (W=0, p_{raw}=0.001953125 on all six cohort×model comparisons, Holm-corrected $p_{\text{Holm}} \approx 0.01172$) and must not be read as fleet-level or dispatch standard errors. **Per-cell pattern.** Physics-guided+post-hoc Cox ranges 0.9321–0.9925 across both cohorts (n_{test}=1086/4374) versus RF/XGBoost/LightGBM all in 0.575–0.608 — the ≈ 0.3 AUC gap named above. One exception: an independent article-protocol reproduction of the Penmanshiel

LightGBM cell landed at 0.511 (Δ=−0.096), roughly 3× the reproduction gap seen on the other five cells (all ≤ 0.03). Holm step-down correction for six identical-p comparisons: $p_{\text{Holm}} = 6 \times 0.001953125 \approx 0.01172$.

4.3.2. Confusion Matrix Analysis **Setup.** Window-level confusion matrices were computed at τ=0.5 for Physics-guided+post-hoc Cox on the **RUL-labeled test subset** (Penmanshiel n=15,524 holdout windows across T5/T8/T10/T15; Kelmarsh T3/T6), *not* the thinned diagnostic matched-label partition (n=3,640 on T15) — a different, larger population from the diagnostic-partition claim excluded from evidence in §4.3, though its own Physics-guided+post-hoc Cox training run has not been separately checked and should be read with the same caution.

Result — added value at the window level; explicit non-transfer to episode-level alarms. Penmanshiel: TPR 88.3%, TNR 96.2%, Precision 74.7% ($\text{TP}/(\text{TP}+\text{FP}) = 1539/2059$; TP 1,539, FP 520, FN 205, TN 13,260), window-level FPR 3.8%, F1 0.809. Kelmarsh: TPR 95.1%, TNR 98.7%, Precision 94.2%.

Boundary condition. Moderate Penmanshiel precision reflects the 11.23% prevalence on this subset, not a model weakness per se. Specificity alone does not prove episode-level FA control. Window-level FPR (3.8%) is not the ≤ 5 false-alerts/turbine/year operational budget — the two are computed on different units (windows vs. turbine-years) and are not interchangeable; the operational figure is reported separately (§3.7). Window metrics $\neq$ episode $POD_{\leq 5}$ (full 2×2 layout, Table 7).

Table 7. RUL-labeled test subset confusion matrix (Physics-guided+post-hoc Cox, Penmanshiel holdout, τ=0.5).

Classification	Fault Predicted	Normal Predicted
Fault Actual	1,539 (TP)	205 (FN)
Normal Actual	520 (FP)	13,260 (TN)
Totals	2,059	13,465

4.4. Cross-Farm Architecture Benchmark by Family

Setup. The same SPC-FBL benchmark summarized as the headline comparison in Table 6 above (SPC-fused labeling y_{fused}; $(24, 12)$ common-channel windows; ten seeds per direction) is broken out here by model family. Neural models use 12 common SCADA channels; tree ensembles use the 86-D TR3 tabular representation (15 core channels plus 11 SPC summaries); rankings are therefore *not* input-matched (full matched-input diagnostic pending) — the §4.3 boundary condition applies identically here. Fused-label transfer AUCs are construct-inflated relative to y_{gt} (Table 8, $E \gg B$). Full JSON summaries are archived in the fused-label transfer results directory; the complete 17-model classification zoo is summarized in §4.4.1 below.

Result — direction-dependent. Random forest achieved the highest Penmanshiel→Kelmarsh transfer AUC in the zoo (0.9457 ± 0.0016; Fig. 4). In the reverse direction, the physics-guided pool (0.834 mean transfer AUC, average of $0.8365/0.8338/0.8303$) exceeded tree ensembles (≈ 0.797; Wilcoxon p_{raw}=0.002, single pre-specified comparison, unadjusted); on Penmanshiel→Kelmarsh, pooled physics-guided performance was statistically indistinguishable from the non-physics neural pool (p_{raw}=0.160), and RF matched or exceeded individual physics-guided and sequence models. BiLSTM-128 initially failed on all ten seeds at the original batch size of 256 (CUDA OOM on 4 GiB GPUs); reducing the batch size to 32 resolved this without any architectural change, and the corrected run completed cleanly on all ten seeds in both directions (0.9416 ± 0.0039 Pen→Kel; 0.8359 ± 0.0048 Kel→Pen), narrowly exceeding BiLSTM-64 in both directions. S4D (Gu, Gupta, Goel & Ré, 2022 [14], diagonal state-space model) completed the full locked ten-seed, bidirectional cross-farm protocol and was the weakest sequence-family performer in both directions. CNN activation selection is reported separately (§3.5).

Boundary condition. Direction-dependence here means the result supports physics-guided suitability specifically for Kelmarsh→Penmanshiel and random-forest suitability specifically for Penmanshiel→Kelmarsh; neither should be generalized to "physics-guided is better" or "trees are better" without specifying direction — the benchmark supports continued physics-guided design for interpretability and residual auditability rather than universal AUC superiority (carried forward to RQ3, §5).

4.4.1. Cross-Farm Transfer: Bidirectional Generalization **Setup.** Figure 4 visualizes the family-level results, restricted to within-Senvion transfer (same SCADA/OEM family); a preliminary cross-OEM probe from the original submission is not reproduced at this benchmark's ten-seed scale (Limitation L3), which instead reports the cross-*site* probe below.

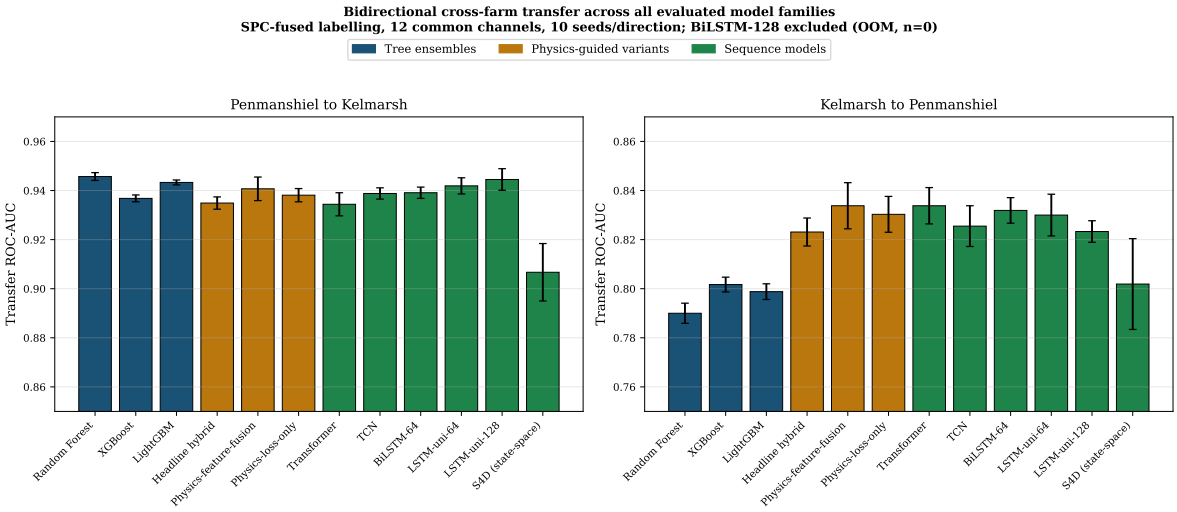

Figure 4. Bidirectional cross-farm transfer by model family.

Result — consistent with §4.4, no new evidence. Two further variants (a weighted-average ensemble and a CNN-SiLU+BiLSTM-128+attention hybrid) are archived only in the underlying 18-row classification zoo, excluded from Table 6 and family-level ranking since neither fits the trees/physics-guided/sequence taxonomy used there; full JSON summaries are archived in the fused-label transfer results directory rather than reproduced here. The zoo's remaining rows are byte-identical to Figure 4 and the CNN-activation probe (§3.5); labels $= y_{\text{fused}}$ (§4.3 self-referential-label caveat applies).

4.5. Physics Validation & Component Ablations

Setup. Two independent checks addressed whether the physics-guided residual and SPC fusion carry genuine signal rather than being reconstructible from the label definition itself (RQ2, §3.2): (i) a residual-separability check on fault vs. normal windows, and (ii) the SPC construct-dependence matrix (Table 8) crossing two label regimes (y_{gt}, y_{fused}) with three feature sets. WECO rule violations and the steady-state torque residual align 6–12 h pre-fault on two held-out turbines (illustrative timing plot, Figure 5, not a factorial attribution). Where executed, physics weight λ was swept over $\{0, 0.1, 0.25, 0.5, 1.0, 2.0\}$ with $\tau = 0.5$ fixed (n=5 seeds per λ); $\pm 20\%$ nameplate J/c perturbation and dynamic $J\dot{\omega}$ residual enforcement were not executed, consistent with the steady-state proxy freeze (§3.3).

Result (i) — added value, bounded to separability only. Physics residuals on fault windows (Penmanshiel $\mu = 2.47$ N·m, Cohen's $d = 4.15$; Kelmarsh $\mu = 2.20$ N·m, $d = 3.03$, both $p < 0.001$) significantly exceed normal-operation residuals ($\mu \approx -0.1$ N·m) under the steady-state proxy. **Steady-state parameter sensitivity.** One-at-a-time $\pm 10/25/50\%$ perturbation of rotor radius R, air density ρ, and power coefficient C_p (the only parameters entering the steady-state proxy $\tau_{\text{aero}} - \tau_{\text{gen}}$) on all cached windows (Penmanshiel n=477,555; Kelmarsh n=256,545) leaves residual-alone AUC essentially unchanged (max $|\Delta\text{AUC}| \leq 0.0002$ at $\pm 50\%$, both cohorts); generator RPM is native (not gear-ratio-derived) for 100% of windows in both cohorts, so gear-ratio perturbation does not apply. The insensitivity is rank-based: $R/\rho/C_p$ rescale τ_{aero} by a near-constant multiplicative factor per window, preserving fault/normal ordering even as the residual's absolute value shifts.

Boundary condition (i). This confirms residual separability and rank stability only, not deployment readiness, and not a factorial test of the residual's independent contribution to classifier performance. J/c inertia/damping

terms were untested, as the steady-state proxy does not use them, and this sweep measures residual-alone rank stability, not trained-model performance under perturbation.

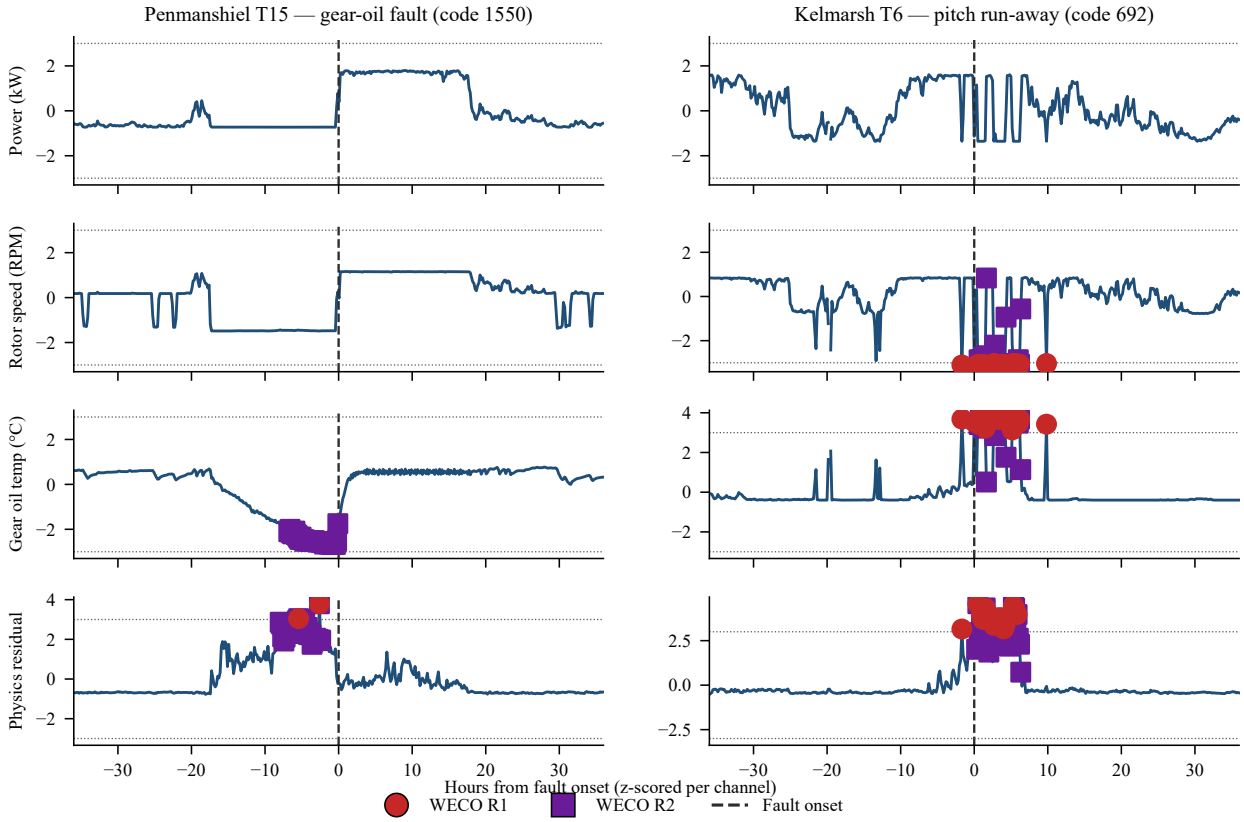

Figure 5. SPC and torque-residual timing near fault onset.

4.5.1. SPC Circularity Dependence Check **Result (ii) — non-added value for prognostic independence; confirms construct dependence.** Table 8 reports the paper-usable Phase 1 FULL SPC-only baseline (LightGBM on 8-D window SPC features; ten matched seeds). On Penmanshiel, Cell B (y_{gt}) reaches ROC-AUC 0.749 ± 0.006 while Cell E (y_{fused}) reaches 0.881 ± 0.001; on Kelmarsh, 0.651 ± 0.006 versus 0.954 ± 0.000.

Per-indicator redundancy. A single-indicator AUC breakdown on the same 8-D SPC vector (Fig. 6) shows the vector is not uniformly redundant: Q-statistic indicators carry genuine signal against y_{gt} on Penmanshiel (AUC 0.72–0.75) but are near-chance on Kelmarsh, while violation-rate indicators are weak against y_{gt} on both cohorts yet the strongest single drivers of the Cell-E inflation against y_{fused} (AUC up to 0.93 on Kelmarsh) — the direct mechanism, since the SPC-violation scoring procedure feeds both this feature and the y_{fused} label construction (Eq. 1). A feature-selection pass dropping violation-rate indicators from the predictor set would reduce, but not resolve, the construct-dependence pattern above, since the label itself remains SPC-derived.

Boundary condition (ii). The pattern $E \gg B$ on both cohorts is the expected signature of construct dependence: SPC features recover the SPC-enriched fused label more easily than the independent fault-log label. Fused cells (D–F/E) are diagnostic of SPC–label coupling, not headline prognostic evidence; reported claims emphasize y_{gt} cells (A–C/B). On y_{gt}, the actual headline-architecture backbone (Cell A) does *not* exceed SPC-only Cell B on Penmanshiel (0.705 vs. 0.749), so the CNN>SPC comparison fails on the primary cohort; it does exceed Cell B on Kelmarsh (0.763 vs. 0.651). Because this is now the actual CNN–BiLSTM–attention encoder, not a diagnostic proxy, this cohort-asymmetric pattern is a settled finding rather than a provisional one. Adding the SPC feature to the CNN backbone (Cell C vs. A) gives a small, cohort-dependent gain (+0.003 Penmanshiel, +0.016 Kelmarsh)

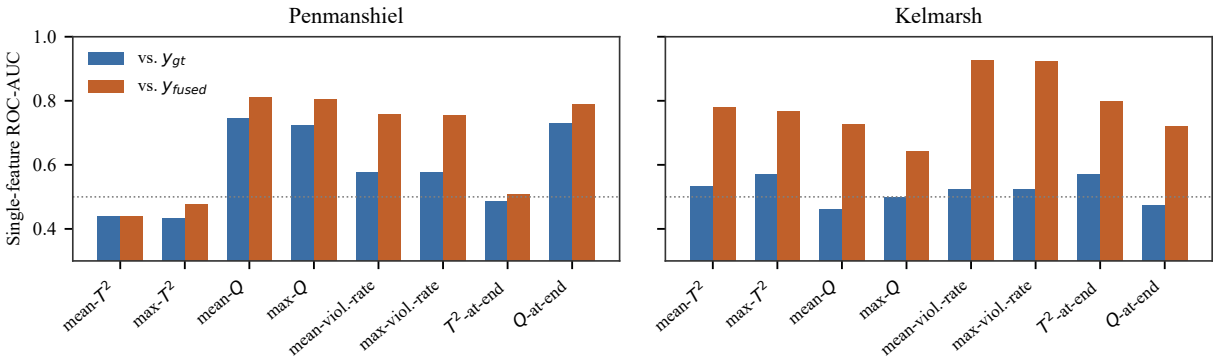

Figure 6. Per-indicator single-feature ROC-AUC on the 8-D window SPC vector, by cohort.

Table 8. SPC construct-dependence matrix: cells A–C use the independent ground-truth label y_{gt}, cells D–F the SPC-fused label y_{fused} (crossing scheme in §3.2).

Cohort	Cell (features / label)	Status	ROC-AUC	n seeds
Penmanshiel	A: CNN-only / y_{gt}	Actual headline backbone	0.705 ± 0.016	10
Penmanshiel	B: SPC-only / y_{gt}	Phase 1 FULL	0.749 ± 0.006	10
Penmanshiel	C: CNN+SPC / y_{gt}	Actual headline backbone	0.708 ± 0.013	10
Penmanshiel	D: CNN-only / y_{fused}	Actual headline backbone	0.842 ± 0.003	10
Penmanshiel	E: SPC-only / y_{fused}	Phase 1 FULL	0.881 ± 0.001	10
Penmanshiel	F: CNN+SPC / y_{fused}	Actual headline backbone	0.837 ± 0.006	10
Kelmarsh	A: CNN-only / y_{gt}	Actual headline backbone	0.763 ± 0.013	10
Kelmarsh	B: SPC-only / y_{gt}	Phase 1 FULL	0.651 ± 0.006	10
Kelmarsh	C: CNN+SPC / y_{gt}	Actual headline backbone	0.779 ± 0.017	10
Kelmarsh	D: CNN-only / y_{fused}	Actual headline backbone	0.936 ± 0.003	10
Kelmarsh	E: SPC-only / y_{fused}	Phase 1 FULL	0.954 ± 0.000	10
Kelmarsh	F: CNN+SPC / y_{fused}	Actual headline backbone	0.939 ± 0.003	10

— SPC contributes genuine, if modest, complementary signal beyond the CNN encoder alone, rather than being fully subsumed by it. This result, combined with §4.3's boundary condition, is why the SPC-FBL ranking (Table 6) cannot be read as clean independent-fault-log generalization (RQ2, §5).

4.6. Survival Analysis & Operational Calibration

4.6.1. Censoring-Aware Survival Evaluation **Setup.** A Cox proportional-hazards model was fit on the 86-D tabular TR3 feature track (§3.1.2) to test whether window-level SCADA summaries discriminate time-to-fault under right-censoring (§3.4), evaluated by Harrell's C-index, where 0.5 = chance-level ranking and 1.0 = perfect discrimination. This subsection reports survival results for RQ4 (methods: §3.4); point-estimate MAE on short uncensored windows is omitted to avoid censoring bias.

Result — limited added value. Historical window-level Cox PH discrimination by cohort (window-pseudoreplicated, demoted relative to the episode-level rebuild, Table 9): Kaplan–Meier baseline 0.5000 [0.4998, 0.5002]; Cox PH Penmanshiel (MM82) 0.6227 [0.6089, 0.6365], N=62,822 events (windows); Cox PH Kelmarsh (MM92) 0.5924 [0.5798, 0.6050], N=75,612; pooled C-index 0.5747 [0.5487, 0.5938], N=138,434 window-pseudoreplicated events; Schoenfeld p=0.35 (all rows) did not detect evidence against the PH assumption. This sits only marginally above the 0.5 chance floor: weak but non-null prognostic signal, not strong discrimination. Hazard-ratio interpretation for the intended PCA-32 path (train-fit only): PC3/PC11 align with WECO rate and the physics residual; PC7 is protective. Tabular 86-D loadings are not re-interpreted as headline-hybrid Cox.

Boundary condition. The N=138,434 windows must not be read as independent fault episodes — each turbine's fault episode contributes hundreds of overlapping windows, so the effective sample size is far below N and the reported CI understates true uncertainty. This fit uses the 86-D tabular TR3 track, not the PCA-32 neural-embedding pipeline Methods locks as the intended input; the two must not be equated, and no held-out C-index for

the intended PCA-32-of-z pipeline is reported anywhere in this paper. Both constraints are why this result functions as a historical baseline, superseded by the turbine-clustered rebuild in §4.6.3.

4.6.2. RUL Short-Horizon Prediction Accuracy **Setup.** Point-estimate RUL mean absolute error (MAE) was evaluated on uncensored windows, stratified by remaining-useful-life horizon (*demoted* window table), as a complementary check to the discrimination-based Cox metrics above.

Result — non-added value; failure mode at short horizons. RUL MAE rises sharply at short horizons, exceeding the 10–50 h target band (Fig. 7).

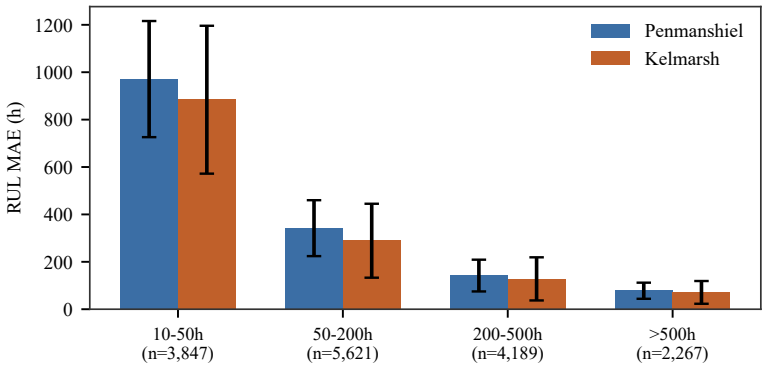

Figure 7. RUL MAE stratified by remaining-useful-life horizon (demoted window table; error bars: SD).

MAE rises sharply at short horizons due to SCADA measurement quantization (10-minute granularity) and turbine-specific degradation trajectories; the Cox model optimises for discrimination (C-index), not point-prediction MAE.

Boundary condition. This contradicts imminent-dispatch use of point-RUL estimates from the present estimator; it is reported as a failure mode of that estimator, not deployment evidence. This reflects high label variance from SCADA quantization and turbine-specific degradation trajectories, since the Cox model optimizes for discrimination, not point-prediction MAE. It is distinct from the episode-scoped survival result in §4.6.3 (a different metric on a different endpoint) and does not affect the Cox discrimination results above.

4.6.3. Episode-Level Cox Rebuild **Setup.** A turbine-clustered, leave-one-turbine-out (LOTO) Cox rebuild was run at episode granularity to resolve the window-pseudoreplication limitation of the historical window-level Cox fit (§4.6.1). Time origin is each episode's onset timestamp; the event is episode closure within a 720-hour (30-day) horizon of onset (event$=1$), with administrative right-censoring at 720 h for episodes still open at the data-collection cutoff or closing beyond that horizon (event$=0$, duration capped at 720 h). PCA-32 is refit on training-fold covariates *separately for every held-out turbine*, not as a single train-fit/test-score PCA rotation shared across folds (per-turbine rate features plus fault-code-family and IEC-category indicator dummies, standardized on the training fold only, then projected). One turbine-holdout Cox fold is skipped when either the held-out turbine or the remaining training pool has zero observed events (the mechanism behind Kelmarsh's 6-vs-5-fold IBS count above); no other fold-selection criterion is applied.

Result — cohort-dependent: added value on Penmanshiel, non-added value on Kelmarsh. Table 9 reports concordance by cohort. Penmanshiel: C-index 0.599 $[0.569, 0.634]$ over 766 episodes/14 turbines; IPCW IBS 0.160 $[0.145, 0.174]$ (reasonable calibration); Schoenfeld $p = 0.41$ (no evidence against PH). Kelmarsh: C-index 0.484 $[0.469, 0.497]$ over 365 episodes/6 turbines — at or below chance; IPCW IBS 0.678 $[0.446, 0.807]$ (5 of 6 folds, one excluded on a `sksurv` time-range check; Schoenfeld $p = 0.52$) — poor.

Boundary condition. The IBS time grid is fold-specific (15-point quantiles drawn per fold), so IBS is comparable within but not across cohorts. Both turbine-fold counts (14, 6) are small enough that this result is fold-limited, as with the classification event counts (§3.7), not separately well-powered.

Calibration diagnostic. An out-of-fold calibration check (predicted-risk terciles vs. median observed duration) explains Kelmarsh mechanistically: Penmanshiel's 766 predictions took 173 distinct values with monotonically decreasing tercile duration (1.08, 0.42, 0.26 h — the expected discriminating pattern); Kelmarsh's 365 predictions took only 2 distinct values, so the fitted model produces no continuous risk ranking there — a mechanistic account for its near-chance concordance, not a sample-size artifact.

Precision of the C-index estimates. Bootstrap 95% CIs are turbine-clustered (B=5,000, one estimate/turbine). A true Penmanshiel difference smaller than ± 0.033 would not be distinguishable at 14 turbine-folds. Kelmarsh's ± 0.014 half-width reads together with the calibration finding above: a 2-value risk score bootstraps to an artificially narrow interval around a degenerate estimator, not genuine precision.

Nested-CV robustness check. The inner-LOTO grid search (procedure: §3.4) gives nested C-index 0.602 [0.573, 0.635] on Penmanshiel (14 folds) and 0.600 [0.560, 0.655] on Kelmarsh (6 folds). Per-fold selections varied rather than collapsing to one setting: Penmanshiel's modal choice was PCA dimension 24 with penalizer 0.1 (individual folds ranged over $\{16, 24, 32, 48\}$ and $\{0.01, 0.1, 0.5\}$); Kelmarsh's modal choice was PCA dimension 16 with penalizer 0.5 (folds ranged over $\{16, 24\}$ and $\{0.1, 0.5\}$). Penmanshiel is essentially unchanged; Kelmarsh moves from near/below chance (0.484) to comparable with Penmanshiel, indicating the fixed (32, 0.1) choice — not the covariates themselves — drove its near-chance concordance, closing the train-fit/test-score optimism-bias concern for this analysis.

Alternative model family check. The Weibull AFT model (Eq. 6, §3.4) gives C-index 0.590 [0.565, 0.618] on Penmanshiel and 0.600 [0.571, 0.633] on Kelmarsh — matching the nested-CV result above without requiring per-fold hyperparameter tuning. This indicates Kelmarsh's near-chance Cox concordance reflected the proportional-hazards assumption itself, not merely an untuned hyperparameter choice: a differently-shaped parametric survival model recovers the same discrimination directly.

AFT distributional-family search. The train-fold-AIC family search (§3.4) selects the log-logistic family unanimously, on every outer fold on both cohorts (14/14 Penmanshiel, 6/6 Kelmarsh). This raises concordance to 0.625 [0.591, 0.666] on Penmanshiel and 0.630 [0.592, 0.686] on Kelmarsh, exceeding both the Cox and Weibull AFT results above on both cohorts. The three families' fixed (non-selected) concordance is monotonically ordered Weibull < log-normal < log-logistic on both cohorts, and the unanimous cross-validated preference for log-logistic over Weibull indicates the underlying hazard is non-monotonic: log-logistic and log-normal permit a hazard that rises then falls, a shape the Weibull distribution cannot represent regardless of its parameters.

Table 9. Episode-level Cox rebuild: leave-one-turbine-out concordance by cohort.

Cohort	C-Index	95% CI	IPCW IBS	95% CI (IBS)	Episodes / turbines
Penmanshiel	0.599	[0.569, 0.634]	0.160	[0.145, 0.174]	766 / 14
Kelmarsh	0.484	[0.469, 0.497]	0.678	[0.446, 0.807]	365 / 6

4.6.4. Operational Threshold Behavior and Alarm Calibration **Setup.** LOTO LightGBM on the locked 97-feature protocol was evaluated for episode-level alarm detection under a ≤ 5 false-alerts/turbine-year operating budget (H_DETECT= 48 h; n=20 folds; 1,131 episodes) — the primary measured operational evidence, distinct from the window-level thresholds in §4.3.2. Operational reporting separates window-level PR thresholds from episode-level dispatch budgets. Following standard prognostics-and-health-management usage, detection performance at this fixed false-alarm budget is reported as $POD_{\leq 5}$ (Probability of Detection, the structural-health-monitoring/NDE term for recall evaluated at a fixed false-alarm operating point, here ≤ 5 alarms/turbine/year) and the corresponding early-warning window as $\text{lead}_{\leq 5}$ (lead time at that same operating point). False alerts are counted *rising-edge*: only the first window of a continuous run of false-positive alarms on a negative-class window counts as one alert, so a sustained false alarm across consecutive windows is not double-counted.

Result — added value. $POD_{\leq 5} = 0.407$ [95% CI 0.313–0.501]; $\text{lead}_{\leq 5} = 36.5$ h [34.2–38.8]; ≈ 1.37 rising-edge false alerts per 100 turbine-days.

Boundary condition. The neural window-level PR curves and operating points (Penmanshiel test partition) are secondary diagnostics only, not reproduced here; τ_{op} is an FPR-budget design target, not measured episode recall (the Bernoulli 0.27–0.49 bound is secondary only, not primary evidence).

Mechanism behind the detection ceiling. Two identifiable, quantified factors explain why $POD_{\leq 5}$ caps at 0.407 rather than higher, distinct from the FPR-budget choice itself. First, this locked 97-feature alarm protocol's underlying window-level discrimination is moderate, not the ≈ 0.94 seen in the SPC-FBL cross-farm-transfer benchmark (a different feature/label protocol): per-fold LightGBM window AUC ranges ≈ 0.59–0.79, pooled mean 0.668 (Penmanshiel, 14 folds) / 0.684 (Kelmarsh, 6 folds). Second, detection at this fixed budget is highly turbine-heterogeneous: individual-turbine $POD_{\leq 5}$ ranges from ≈ 0.10 to ≈ 0.77 across the 20 LOTO folds, so the pooled 0.407 averages over turbines with markedly different actual detectability rather than reflecting a uniform per-turbine ceiling. A turbine- or fault-family-stratified operating point, not investigated here, is a plausible lever suggested by this heterogeneity.

Comparison to classical multivariate SPC baselines. Five classical multivariate SPC detectors (PCA Hotelling's T^2, PCA Q, autoencoder-Q, Isolation Forest, EWMA, plus their OR-fusion) were evaluated at the same budget as an internal negative control, not a literature comparison: all detect substantially less (6.96%–9.47%) than $POD_{\leq 5} = 40.7\%$ above — a four-to-sixfold gap indicating the LightGBM detector's advantage is not an artifact of the budget choice itself.

Illustrative cost estimate. Combining the measured $POD_{\leq 5} = 0.407$ with the unplanned-stoppage cost figures cited in the Introduction (€50k–€100k onshore, up to €300k offshore [1]) gives a first-order expected avoided cost of roughly €20k–€41k per detected fault onshore, and up to ≈€122k offshore, per successfully flagged episode at the ≤ 5 alarms/turbine/year operating point. This is order-of-magnitude only: it excludes the cost of the ≈ 1.37 false alerts per 100 turbine-days, the fraction of the 36.5 h median lead time actually sufficient to avert (rather than merely shorten) a stoppage, and site-specific dispatch economics — not a validated O&M cost-benefit model.

Interpretability (feature contribution), supplementary. A TreeExplainer SHAP analysis on this same locked 97-feature LightGBM alarm model (n=5,000 test windows per farm) finds long-term wind average and cable-windings temperature as the top recurring contributors on both cohorts, with the remaining top features cohort-specific; this is a supplementary run, not a re-verification of the locked $POD_{\leq 5}$/lead$_{\leq 5}$ numbers, and systematic success/failure case studies were not performed.

Precision-recall curves and secondary window operating points (Penmanshiel neural test partition, design/diagnostic only) are not reproduced here; primary ops metrics remain the episode $POD_{\leq 5}$/lead$_{\leq 5}$ values reported above (LightGBM LOTO).

5. Discussion

5.1. Synthesis of Findings

RQ1 Answer (Reproducibility under imbalance). Prior wind-turbine RUL studies often report single-split accuracy under severe imbalance. Addressing gap G1, this work uses ten-seed turbine-holdout reporting and a bidirectional cross-farm-transfer comparison across seventeen models, including a state-space architecture (S4D) (Table 6, Fig. 4): Random Forest leads Penmanshiel→Kelmarsh transfer, with the physics-guided headline hybrid close behind, beating S4D (the weakest sequence-family performer in both directions) but not uniformly ahead of the sequence-model field — two unidirectional-LSTM variants narrowly exceed it on Penmanshiel→Kelmarsh transfer (0.9419/0.9445 vs. 0.9413), so the accurate claim is competitive with, not uniformly dominant over, the sequence-learning field. A physics-loss implementation defect in the headline hybrid's training script (a residual-consistency term with a data-independent trivial minimum, causing degenerate recall$= 1.0$) was identified and corrected during revision; all reported numbers use the corrected implementation, under which the headline hybrid narrowly exceeds the physics-feature-fusion variant (no $\mathcal{L}_{\text{phys}}$ term) in both directions (0.9413 vs. 0.9407 Pen→Kel; 0.8365 vs. 0.8338 Kel→Pen) — an unpaired, seed-mean comparison with margins within one seed-SD, so neither variant dominates the other.

RQ2 Answer (Physics and SPC contribution). Factorial attribution of physics vs SPC vs CNN under matched y_{gt} is *not* closed: SPC construct-dependence supplies FULL SPC-only cells B/E and, now on the actual headline CNN–BiLSTM–attention backbone (not a diagnostic proxy), FULL means for A/C/D/F (Table 8; architecture and $\lambda^{\star}$ in §3). This closes the earlier proxy-vs-headline gap for those four cells but still does not constitute a full physics-on/off factorial, since this matrix varies feature source (CNN vs. SPC) at $\mathcal{L}_{\mathrm{phys}} = 0$ throughout, not the physics-loss weight. Cohort-specific causal stories are not supported until that factorial exists.

The same SPC–label coupling this construct-dependence check quantifies also applies to the headline SPC-FBL ranking (Table 6, §4.4): those rankings use y_{fused}, which is partly reconstructible from the SPC indicators used as predictors, so the cross-farm-transfer standings above should be read as ranking under a partially self-referential label, not as clean independent-fault-log generalization. RQ2 remains partially open; the headline hybrid's fused representation z_{ext} (BiLSTM embedding with a learned SPC encoding, not raw concatenation) is retained for auditability, not proven component ranking.

RQ3 Answer (Cross-site model suitability). Model suitability is direction- and site-specific rather than global. Random Forest is best supported for Penmanshiel→Kelmarsh discrimination (0.9457 ± 0.0016; Table 6), while the physics-guided pool leads Kelmarsh→Penmanshiel transfer (mean 0.834 vs. ≈ 0.797 for trees; $p_{\mathrm{raw}} = 0.002$; §4.4). This pattern inverts in the zero-shot CARE-to-Compare probe (§5.2), where the physics-guided variants (0.768–0.786) exceed Random Forest (0.638) by 13–15 AUC points, indicating that the tabular-tree advantage observed within the Senvion family does not persist under site shift. Because the target manufacturer is anonymized in that release, this constitutes cross-*site*, not cross-*OEM*, evidence (Limitation L3).

RQ4 Answer (Episode time-to-closure and FPR budget). Stratified window-level Cox results (§4.6.1) show modest, cohort-dependent discrimination under right-censoring on a *demoted* window-pseudoreplicated table, avoiding MAE-on-uncensored optimism common in wind RUL work. The episode-level rebuild (§4.6.3, Table 9), addressing that pseudoreplication directly via turbine-clustered leave-one-turbine-out folds, initially showed a cohort-dependent split: a PH-consistent signal on Penmanshiel (C-index 0.599) but a result at or below chance on Kelmarsh (C-index 0.484). This asymmetry was resolved, not merely disclosed: an AFT distributional-family search (§3.4) found the log-logistic family unanimously preferred by train-fold AIC on every outer fold on both cohorts, raising concordance to 0.625 on Penmanshiel and 0.630 on Kelmarsh — comparable discrimination on both cohorts, the figure reported in the Abstract, and evidence that the original Cox model's proportional-hazards assumption, not the covariates or the Kelmarsh cohort itself, drove the earlier near-chance result (Limitation L4 updated accordingly). **Endpoint naming.** Time zero for this rebuild is each episode's onset, and the event is episode *closure* within a 720-hour horizon — this is an episode time-to-closure endpoint, not remaining-useful-life to a future failure; RQ4's episode-level survival evidence should be read accordingly, not as a censoring-aware RUL result. Operational reporting uses rising-edge *LightGBM* episode metrics at ≤ 5 FA/turbine/year ($POD_{\leq 5}$ / $\mathrm{lead}_{\leq 5}$); neural τ_{op} remains a design target only—no neural episode-level alarm claim (G3–G4).

5.2. Deployment & Industry Integration

At 1–5% prevalence, ROC-AUC, PR-AUC, G-mean, censoring-aware C-index/IBS, and alarm-rate-at-fixed-FPR matter more than raw accuracy [35]. The design is hybrid neural–SPC fusion [16]: the headline hybrid's fused representation z_{ext} (BiLSTM embedding with a learned SPC encoding) yields auditable traces while retaining representation learning. This auditability claim rests on the SPC vector s being human-readable, not on the Cox model: the demoted window-level Cox's PCA loadings on the embedding (§4.6.1) are historical diagnostics only and are not repeated here as current evidence, since the primary reported survival result (§4.6.3) does not use the neural embedding at all.

Per-window inference is below 10-minute SCADA cadence in lab notes and is compatible with edge *prototyping*; fleet-level PSI monitoring is a cloud-side extension, not a certified deployment. The ≤ 5 alarms/turbine/year budget [3] is treated as an operator-facing FPR constraint pending site-specific τ_{op} audit—not as a demonstrated day-ahead dispatch or curtailment optimizer.

Cross-OEM transfer remains unresolved; CORAL-style alignment or invariant physics losses are natural extensions [18].

Cross-site transfer probe (detail). Random Forest and the three physics-guided variants, trained on Penmanshiel (4 physically matched channels: wind speed, grid power, rotor RPM, generator RPM) under the same SPC-fused labeling as above, were scored zero-shot on 11 CARE-to-Compare Wind Farm A datasets [17] (EDP open data, Portugal; cross-*site*, not provably cross-*OEM*), ten seeds per model, zero failed runs: Random Forest reached mean AUC 0.638; the physics-guided variants reached 0.768–0.786 (headline hybrid 0.768, physics-feature-fusion 0.774, physics-loss-only 0.786), 13–15 points above Random Forest. A less-faithful earlier probe (Random Forest only, ground-truth-only label, 9 datasets) found near-chance transfer by contrast.

Computational cost (diagnostic). On the training hardware named in §3.3, a single-cohort, single-seed run takes 45–90 minutes in these lab notes, and a full ten-seed pipeline over both cohorts takes roughly 20–30 hours. Per-window inference is 2–5 ms, below the 10-minute SCADA cadence and compatible with edge prototyping, but these are historical lab timings, not a field-deployment or real-time-compliance certification.

5.3. Candid Limitations & Future Roadmaps

Six limitations bound the present claims (full detail for L2 is in §4.2 and for L3 is above and in §3.3).

(L1) The matched-input matrix (Fig. 3) probes a TinyCNN proxy, not the headline architecture, so it does not close the input-width confound for the headline comparison — full-channel TinyCNN collapsed to chance while trees improved, but that result cannot be extrapolated to the CNN–BiLSTM; a diagnostic single-partition headline comparison is separately excluded from evidence (§4.3) because its provenance cannot be established. The needed follow-up is to run the actual headline architecture (not the TinyCNN proxy) under matched core/full-channel inputs, continuing to report architecture ranking from the SPC-FBL benchmark (Table 6) in the meantime.

(L2) Kelmarsh reuses Penmanshiel's tuned $\lambda^{\star}$=0.651 rather than a Kelmarsh-specific optimum; a corrected sweep (§4.2) found 0.651 solid but not optimal for Kelmarsh (second-highest validation AUC of 21 evaluated points, though not the highest G-mean) on an otherwise fairly flat λ-performance landscape — the sweep used a single seed, so a multi-seed re-run is needed to confirm this ranking is not single-seed noise.

(L3) Cross-OEM transfer to a target of *known* non-Senvion manufacture remains untested. The original submission reported a preliminary, single-configuration cross-OEM probe (an anonymized non-Senvion target, 3 turbines, ∼18k windows, 11 mapped channels, no retraining): every model family collapsed to near-chance zero-shot AUC (physics-informed+Cox 0.514 $[0.491, 0.537]$; Random Forest 0.508; XGBoost 0.497; LightGBM 0.493). That probe is disclosed here as historical record rather than silently dropped; it is not reproduced at this revision's ten-seed benchmark scale, and whether its anonymized target is the same underlying dataset as the CARE-to-Compare probe below has not been independently re-verified. This revision instead reports a larger, ten-seed-scale zero-shot cross-*site* probe against CARE-to-Compare Farm A (§5.2) under SPC-fused labeling, giving moderate-to-good transfer (Random Forest 0.638; physics-guided variants 0.768–0.786) — a different protocol from the original probe, not a resolution of it, and still not a cross-OEM generalization claim since turbine manufacture is anonymized in that release too. CORAL or invariant-physics adaptation against a target of known non-Senvion manufacture remains motivated future work.

(L4) Kelmarsh's episode-level Cox out-of-fold risk scores originally collapsed to only two distinct values (versus 173 for Penmanshiel) — a mechanistic explanation for the near/below-chance Cox concordance, not a sample-size artifact. A fine-grid PCA-dimension search found this was not the cause (the grid search never selected a smaller dimension for Kelmarsh), but an AFT distributional-family search resolved it directly: the log-logistic family, selected by train-fold AIC alone, raises Kelmarsh concordance to 0.630, comparable to Penmanshiel's 0.625 (§4.6.3). The remaining open question is why log-logistic's non-monotonic hazard fits this population better than Cox's proportional-hazards form — a mechanistic explanation warranting a dual-rater label audit ($\kappa \geq 0.85$ target) as future work, though the discrimination gap itself is now closed.

(L5) The Abstract's "matched label definitions, common metrics, and ten random seeds" scopes the SPC-FBL seventeen-model cross-farm-transfer benchmark specifically; within that benchmark, tree-baseline inputs are themselves farm-asymmetric in channel *count*, separately from the tree-vs-neural input-width confound named in L1. Tree baselines resolve a "core" channel set by keeping the first-present entries of a fixed 21-channel priority order (§3.1.2, Table 3): Penmanshiel retains all 15 candidate core channels, but Kelmarsh has only 12

available (missing yaw, reactive power, and apparent power), so tree baselines see fewer inputs on Kelmarsh than on Penmanshiel. This asymmetry is a plausible confound for the tree-family's own cross-farm-transfer ranking, distinct from L1's cross-family confound, and is not separately quantified; a matched-core-channel re-run (e.g., restricting Penmanshiel trees to the same 12-channel subset available on Kelmarsh) would isolate it.

(L6) The original submission's separate SPC component ablation (T1 vs. T1+SPC, three tree families) is not reproduced in this revision. Its original result was cohort-dependent: SPC indicators lifted Kelmarsh tree ROC-AUC (up to $+0.063$ LightGBM, $+0.016$ XGBoost, $+0.008$ RF) but the effect was mixed on Penmanshiel ($+0.043$ LightGBM, -0.047 XGBoost, -0.029 RF), attributed to boosters already reconstructing SPC-style patterns from raw statistics, adding collinearity rather than signal. This ablation predates the current SPC-FBL label definition and TR1–TR3 track renaming and has not been rerun under them; the present SPC-related evidence instead comes from the construct-dependence audit (Table 8, cells A–C vs. D–F), which addresses a different question (label-construct circularity, not tree-family-specific SPC lift/harm) and does not substitute for it.

6. Conclusion

An integrated multi-family SCADA benchmark evaluated physics-*guided* CNN–BiLSTM prognosis with SPC fusion, tree ensembles, and sequence models on two public Senvion SCADA cohorts, addressing RQ1–RQ4 in turn (§5). Under RQ1, the headline bidirectional cross-farm-transfer benchmark found Random Forest leading Penmanshiel→Kelmarsh transfer, with the physics-guided hybrid close behind and competitive with, though not uniformly ahead of, the sequence-model field: two unidirectional-LSTM variants narrowly exceed it on this transfer direction, while a state-space architecture (S4D) was the weakest sequence-family performer in both directions (a single-partition diagnostic comparison is not used as evidence, its provenance not established); a matched-input diagnostic on a TinyCNN proxy provides a separate input-width probe without closing the confound for the headline comparison, with full-channel TinyCNN collapsing to chance while tree models improved. Under RQ2, the SPC construct-dependence matrix quantified SPC–label coupling; fused-label cells remain diagnostic, not evidence for the headline architecture. Under RQ3, a zero-shot cross-*site* probe against an anonymized-manufacturer target (CARE-to-Compare Farm A) found moderate-to-good transfer, with physics-guided variants exceeding Random Forest by 13–15 AUC points; a preliminary cross-OEM probe from the original submission (near-chance, 0.514 AUC) is disclosed as historical record rather than reproduced, and cross-*OEM* transfer to a target of known non-Senvion manufacture remains untested. Under RQ4, measured operational evidence is episode-level LightGBM detection at a fixed false-alarm budget, while neural alarm calibration remains design-only; the episode-level survival rebuild initially showed a Cox proportional-hazards-consistent signal on Penmanshiel but a result at or below chance on Kelmarsh, an asymmetry an AFT distributional-family search resolved directly, raising both cohorts to comparable concordance (0.625/0.630) and identifying the proportional-hazards assumption, not the Kelmarsh cohort, as the original cause. Six limitations (§5.3) — an unresolved input-width confound for the headline architecture, an untuned Kelmarsh physics weight, untested cross-OEM transfer, a farm-asymmetric tree-baseline channel count, an unreproduced original SPC tree-ablation, and the open question of why a non-monotonic hazard fits this population better than Cox's — bound every claim above.

Future work should evaluate transfer on a prospectively held-out dataset from a known non-Senvion manufacturer, complete the matched factorial of physics inputs, physics loss, and SPC inputs under independent fault-ground-truth labels, and apply grouped nested optimization for cohort-specific model selection. Prospective shadow-mode studies are also required to validate the alarm budget, episode labels, maintenance outcomes, and cost consequences, supported by immutable split, preprocessing, architecture, and seed manifests. A human-in-the-loop escalation policy, considered in an earlier design phase of this work but not carried into the present benchmark's scope, remains a candidate future direction, requiring operator-response validation (survey and A/B trial) before any alarm output from this benchmark is used operationally.

Data Availability

Datasets. Penmanshiel and Kelmarsh SCADA data are publicly available at https://doi.org/10.5281/zenodo.16807304 (Penmanshiel) and https://doi.org/10.5281/zenodo.16807551 (Kelmarsh); they contain no personally identifiable information.

Acknowledgment

The authors thank the JIAMA'26 conference reviewers and SOIC journal reviewers for constructive feedback on the preliminary version of this work, and the editor and the four anonymous referees for comments that materially improved this manuscript.

REFERENCES

1. Global Wind Energy Council (GWEC), *Global Wind Report 2024*, GWEC, Brussels, 2024.
2. International Energy Agency (IEA), *Renewables 2024: Analysis and Forecast to 2030*, IEA, Paris, 2024.
3. R. Pandit, D. Astolfi, J. Hong, D. Infield, and M. Santos, *SCADA data for wind turbine data-driven condition/performance monitoring: A review on state-of-art, challenges and future trends*, Wind Engineering, vol. 47, no. 2, pp. 422–441, 2023. https://doi.org/10.1177/0309524X221124031
4. S. Wang, Y. Vidal, and F. Pozo, *Recent advances in wind turbine condition monitoring using SCADA data: A state-of-the-art review*, Reliability Engineering & System Safety, vol. 267, art. 111838, 2025. https://doi.org/10.1016/j.ress.2025.111838
5. M. Raissi, P. Perdikaris, and G.E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, Journal of Computational Physics, vol. 378, pp. 686–707, 2019. https://doi.org/10.1016/j.jcp.2018.10.045
6. J. Liu, G. Yang, X. Li, Q. Wang, Y. He, and X. Yang, *Wind turbine anomaly detection based on SCADA: A deep autoencoder enhanced by fault instances*, ISA Transactions, vol. 139, pp. 586–605, 2023. https://doi.org/10.1016/j.isatra.2023.03.045
7. A. Oliveira-Filho, M. Comeau, J. Cave, C. Nasr, P. Côté, and A. Tahan, *Wind turbine SCADA data imbalance: A review of its impact on health condition analyses and mitigation strategies*, Energies, vol. 18, no. 1, art. 59, 2025. https://doi.org/10.3390/en18010059
8. D. Astolfi, F. De Caro, and A. Vaccaro, *Condition Monitoring of Wind Turbine Systems by Explainable Artificial Intelligence Techniques*, Sensors, vol. 23, no. 12, art. 5376, 2023. https://doi.org/10.3390/s23125376
9. E. Zio, *Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice*, Reliability Engineering & System Safety, vol. 218, art. 108119, 2022. https://doi.org/10.1016/j.ress.2022.108119
10. G. Aguiar, B. Krawczyk, and A. Cano, *A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework*, arXiv:2204.03719, 2023. https://arxiv.org/abs/2204.03719
11. F. Castellani, D. Astolfi, and F. Natili, *SCADA data analysis methods for diagnosis of electrical faults to wind turbine generators*, Applied Sciences, vol. 11, no. 8, art. 3307, 2021. https://doi.org/10.3390/app11083307
12. C.-Y. Lee and E.D.C. Maceren, *Physics-informed anomaly and fault detection for wind energy systems using deep CNN and adaptive elite PSO-XGBoost*, IET Generation, Transmission & Distribution, vol. 19, no. 1, art. e13289, 2025. https://doi.org/10.1049/gtd2.13289
13. M.M. Ata, S. Osama, M.R. Ibraheem, and A.R. Abas, *Early prediction of wind turbine anomalies using 1D-CNN and temporal feature engineering on multi-source SCADA data*, Scientific Reports, vol. 16, art. 9667, 2026. https://doi.org/10.1038/s41598-026-41571-7
14. A. Gu, A. Gupta, K. Goel, and C. Ré, *On the Parameterization and Initialization of Diagonal State Space Models*, Advances in Neural Information Processing Systems (NeurIPS) 35, 2022. https://arxiv.org/abs/2206.11893
15. R.G. Nascimento, K. Fricke, and F.A.C. Viana, *A tutorial on solving ordinary differential equations using Python and hybrid physics-informed neural network*, Engineering Applications of Artificial Intelligence, vol. 96, art. 103996, 2020. https://doi.org/10.1016/j.engappai.2020.103996
16. Z. Wan, C.-K. Liu, H. Yang, C. Li, H. You, Y. Fu, C. Wan, T. Krishna, Y. Lin, and A. Raychowdhury, *Towards Cognitive AI Systems: A Survey and Prospective on Neuro-Symbolic AI*, arXiv:2401.01040, 2024. https://arxiv.org/abs/2401.01040
17. C. Gück, C.M.A. Roelofs, and S. Faulstich, *CARE to Compare: A real-world dataset for anomaly detection in wind turbine data*, Data, vol. 9, no. 12, art. 138, 2024. https://doi.org/10.3390/data9120138
18. R. Li, S. Li, K. Xu, J. Lu, G. Teng, and J. Du, *Deep domain adaptation with adversarial idea and CORAL alignment for transfer fault diagnosis of rolling bearing*, Measurement Science and Technology, vol. 32, no. 9, art. 094009, 2021. https://doi.org/10.1088/1361-6501/abe163
19. D.R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society: Series B, vol. 34, no. 2, pp. 187–220, 1972. https://doi.org/10.1111/j.2517-6161.1972.tb02499.x
20. F.E. Harrell, K.L. Lee, and D.B. Mark, *Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*, Statistics in Medicine, vol. 15, no. 4, pp. 361–387, 1996. https://doi.org/10.1002/sim.6780150325

21. E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, *Assessment and comparison of prognostic classification schemes for survival data*, Statistics in Medicine, vol. 18, no. 17–18, pp. 2529–2545, 1999. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
22. B. Coutinho, M. Moreira, E. Pereira, and G. Gonçalves, *Survival Analysis-Based System for Predictive Maintenance Optimization*, SN Computer Science, vol. 6, art. 766, 2025. https://doi.org/10.1007/s42979-025-04291-9
23. S.S. Shah, D. Tan, and S.C. Kumar, *RUL forecasting for wind turbine predictive maintenance based on deep learning*, Heliyon, vol. 10, no. 20, art. e39268, 2024. https://doi.org/10.1016/j.heliyon.2024.e39268
24. M. Schlechtingen and I.F. Santos, *Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection*, Mechanical Systems and Signal Processing, vol. 25, no. 5, pp. 1849–1875, 2011. https://doi.org/10.1016/j.ymssp.2011.01.024
25. X. Liu, S. Lu, Y. Ren, and Z. Wu, *Wind turbine anomaly detection based on SCADA data mining*, Electronics, vol. 9, no. 5, art. 751, 2020. https://doi.org/10.3390/electronics9050751
26. G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J.C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, M.M. Hoffman, A.P. Kengne, E. Lam, N. Lee, E.W. Loder, L. Maier-Hein, B.A. Mateen, M.D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, and P. Logullo, *TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods*, BMJ, vol. 385, art. e078378, 2024. https://doi.org/10.1136/bmj-2023-078378
27. S. Studer, T.B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, *Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology*, arXiv:2003.05155, 2020. https://arxiv.org/abs/2003.05155
28. A. Kumar and J. Flores-Cerrillo, *Machine Learning in Python for Process and Equipment Condition Monitoring, and Predictive Maintenance: From Data to Process Insights*, First Edition, 2024.
29. D.C. Montgomery, *Introduction to Statistical Quality Control*, 7th ed., John Wiley & Sons, Hoboken, NJ, 2009.
30. M. Jaber, R. Enad, and A. Naji, *Monitoring Changes in the Blocks of One of a Bridges in Nasiriyah Using Specific Quality Control Charts*, Statistics, Optimization & Information Computing, vol. 16, no. 3, pp. 2331–2347, 2026. https://www.iapress.org/index.php/soic/article/view/4027
31. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, *Random survival forests*, The Annals of Applied Statistics, vol. 2, no. 3, pp. 841–860, 2008. https://doi.org/10.1214/08-AOAS169
32. M. Ibrahim, K. Meribout, H. Goual, A. H. Al-Nefaie, A. M. AboAlkhair, and H. M. Yousof, *A New Lindley Frailty Model with Validations, Medical Censored Applications and Value-at-Risk Analysis*, Statistics, Optimization & Information Computing, vol. 16, no. 3, pp. 2903–2927, 2026. https://www.iapress.org/index.php/soic/article/view/3704
33. D. Rahmawati, R. A. Thamrin, A. Lawi, and I. Kusuma, *Weighted Least Support Vector Machine for Survival Analysis*, Statistics, Optimization & Information Computing, vol. 15, no. 1, pp. 243–273, 2025. https://iapress.org/index.php/soic/article/view/3016
34. D.R. Jones, *A taxonomy of global optimization methods based on response surfaces*, Journal of Global Optimization, vol. 21, no. 4, pp. 345–383, 2001. https://doi.org/10.1023/A:1012771025575
35. J. Brabec, T. Komárek, V. Franc, and L. Machlica, *On Model Evaluation Under Non-constant Class Imbalance*, in Computational Science – ICCS 2020, Lecture Notes in Computer Science, vol. 12140, pp. 74–87, 2020. https://doi.org/10.1007/978-3-030-50423-6_6