A Collective Learning Workflow for Remaining Useful Life Estimation of Building Assets Under Sparse Degradation Data

Edgar Segovia[1], Joao Patacas[1], Xiang Xie[1], Philip James[1], Sneha Verma[1] and Mohamad Kassem[1]

[1]*School of Engineering, Newcastle University, Newcastle, United Kingdom*

Edgar.Segovia@newcastle.ac.uk
Joao.Patacas@newcastle.ac.uk
Xiang.Xie@newcastle.ac.uk
Philip.James@newcastle.ac.uk
Sneha.Verma@newcastle.ac.uk
Mohamad.Kassem@newcastle.ac.uk

ABSTRACT

Building maintenance in the commercial sector is still dominated by reactive and schedule-based strategies, yet the shift to predictive maintenance is held back by a chronic shortage of degradation data: individual assets such as fan coil units, air handling units and pumps rarely accumulate enough failure or degradation history to train standalone prognostic models, and the records that do exist are short and heavily skewed towards normal operation. Existing approaches do not resolve this. Deep sequence models require continuous multi-year recordings that newly instrumented buildings cannot provide and overfit on small, imbalanced datasets; semantic ontologies organise building data but carry no prognostic capability; and baseline forecasting methods generate operating features without estimating degradation. This paper presents a collective-learning workflow that groups functionally identical assets by their Brick semantic class, pools their operational records into a shared dataset, and trains a single Random Forest model to estimate each asset's daily degradation increment, which is then accumulated into a remaining-useful-life projection and updated as maintenance is recorded. Applied to operational data from a pilot building, the workflow produced per-unit daily degradation estimates and remaining-useful-life projections with quantified uncertainty for fan coil units that individually lacked sufficient history to be modelled in isolation. Under leave-one-out cross-validation, operational features alone did not recover the per-unit degradation level (cross-validated R^2 below zero), whereas fusing a static condition indicator with the operational dynamics was required to do so (R^2 = 0.67, CVRMSE 20 percent), reported as a preliminary finding rather than a validated prognostic capability. The results show that pooling sparse degradation labels across semantically aligned assets makes data-driven prognostics feasible for building portfolios under data scarcity, providing a transferable route to predictive maintenance that does not depend on multi-year failure records.

1. INTRODUCTION

Building maintenance in the commercial sector is dominated by reactive and schedule-based strategies that fail to anticipate equipment degradation. Predictive maintenance (PdM) offers a data-driven alternative, but its application to building systems faces a structural obstacle: individual assets such as fan coil units, air handling units, and pumps rarely accumulate sufficient failure or degradation records to train standalone prognostic models (Asare, Liu and Anumba, 2025). Operational data is often limited to months rather than full degradation cycles, failure events are extremely rare, and the available history covers only a fraction of each asset's expected lifespan (Hosseini Gourabpasi and Nik-Bakht, 2024; Hu and Cai, 2024).

Yet building portfolios contain populations of functionally identical assets, the same type of fan coil unit repeated across floors, the same pump model installed in multiple plant rooms that collectively span a wider range of operating conditions than any single unit. This paper exploits this observation through a collective learning approach: grouping assets by their semantic class within a building ontology and pooling their operational data to construct a shared training dataset from which a common degradation model is learned. Throughout this paper, the term collective training data refers specifically to this pooled, semantically aligned operational record. It is a single tabular dataset in which each row is a day. The features describe the daily operating conditions of an individual unit, and the labels are the daily degradation increments derived from the asset health histories of all units of the same Brick class. The

workflow is therefore a mechanism for propagating scarce degradation labels across a fleet of like assets under data scarcity, rather than a model that learns a physical degradation law from operational behaviour alone: the operational features carry the daily dynamics, while a static condition indicator anchors the per-unit degradation level (Section 4).

The proposed workflow assumes that each asset has been pre-classified into a standardised ontological class and that its current health state has been quantified as an Asset Health Index (AHI) on a 0 to 100 scale, derived from the asset's chronological age and its most recent physical condition rating. These inputs serve as the prognostic starting condition. The workflow then proceeds through a four-level architecture, presented in full in Section 3: Level 1 generates climate-matched daily usage baselines; Level 2 estimates the daily degradation increment for each unit with a collectively trained model; Level 3 accumulates these increments to determine the Remaining Useful Life (RUL); and Level 4 retrains the model whenever a maintenance intervention is recorded.

The paper is structured as follows. Section 2 reviews the relevant literature on data scarcity, ensemble methods, baseline forecasting, and ontological interoperability for building PdM, and ends by listing the scientific issues addressed in this work. Section 3 presents the method: the architecture overview, the ontology layer, baseline generation, collective model training, daily degradation estimation, and the predictive-model instantiation. Section 4 reports a preliminary application of the workflow to operational data from a pilot building. Section 5 discusses the implications and limitations, and Section 6 concludes with directions for future work.

2. LITERATURE REVIEW

2.1. Data scarcity in building predictive maintenance

The application of data-driven predictive maintenance to building systems is constrained by a chronic shortage of degradation and failure data. Scant historical data has been identified as a general barrier across the building sector (Asare, Liu and Anumba, 2025), while insufficient fault data has been reported as the primary obstacle to chiller fault detection and diagnostics (Ma et al., 2023). The limited availability and diversity of sensory data in commercial buildings further restricts the development of reliable models (Hosseini Gourabpasi and Nik-Bakht, 2024), and practical field datasets are often highly imbalanced, with far fewer failure samples than normal-operation records (Hu and Cai, 2024). Standard supervised learning approaches tend to overfit to the dominant healthy-operation class when trained on such imbalanced data (Hosamo et al., 2022).

This scarcity is compounded by a non-linearity problem: the rate at which a building asset degrades depends on its usage intensity, external climate, and maintenance quality, not simply on its age. Models trained in controlled environments do not necessarily generalise to real buildings (Hosseini Gourabpasi and Nik-Bakht, 2024), and data-driven models have been shown to lose predictive power when the operating environment changes (Kang, Chung and Hong, 2021). These findings reinforce the need for approaches that account for actual operating conditions rather than static assumptions.

2.2. Deep learning limitations and ensemble alternatives

Deep learning architectures such as Long Short-Term Memory (LSTM) networks are widely used in industrial prognostics because they capture temporal dependencies in sequential data. However, they require long, continuous time-series spanning multiple degradation cycles, which are unavailable in newly commissioned or recently instrumented buildings (Hu et al., 2023). Complex architectures carry high computational cost and a strong tendency to overfit when datasets are small (Hakimi, Liu and Abudayyeh, 2025; Guzmán-Torres et al., 2025), and the scarcity of severe failure events in building datasets has been found to limit the efficacy of LSTM-based failure prediction specifically (Hu et al., 2023).

Ensemble methods such as Random Forest offer a practical alternative. It aggregates predictions from multiple independent decision trees, making it inherently resistant to overfitting with small datasets. A comparative evaluation of ten machine learning algorithms for fault prediction in building HVAC systems found that ensemble methods, including Random Forest, achieved competitive accuracy with significantly lower data requirements than deep learning architectures (Hosamo et al., 2023). Critically, because each daily observation can be treated as an independent tabular sample characterised by operational features, Random Forest does not require the sequential input structure that LSTM demands.

2.3. Baseline methods for building energy and temperature

Forecasting the expected usage of a building asset is a prerequisite for estimating its degradation under future operating conditions. The building energy literature offers several baseline methods for this purpose. The X-of-Y algorithm, a variant of the nearest-neighbour approach, selects historical days with similar outdoor temperatures and averages their energy profiles to predict a target day. This method has been shown to achieve acceptable accuracy (CVRMSE in the range of 10-15%) with minimal computational overhead and no requirement for multi-year training data (Hyndman and Koehler, 2006). ASHRAE Guideline 14 provides standard criteria for baseline model validation, establishing CVRMSE thresholds that are widely used as benchmarks in the field (ASHRAE, 2014).

2.4. Semantic interoperability for collective approaches

A practical prerequisite for pooling operational data across assets is a standardised vocabulary that ensures the data from different buildings refers to the same type of equipment. The Brick schema is an open-source ontology specifically designed for the building domain, providing a hierarchical vocabulary for equipment, sensors, spatial elements, and their relationships (Balaji et al., 2018). It has been adopted in several digital twin and smart building research projects as the standard for equipment classification (Hosamo et al., 2023). By classifying assets into Brick classes before pooling their data, a collective training approach can ensure that the aggregated dataset contains only operationally homogeneous samples, which is a necessary condition for the shared model to generalise across individual units.

2.5. The collective training gap

Despite these advances, a structural gap remains: no existing method combines semantic asset grouping, climate-based usage baselines, and collectively trained machine-learning models into an integrated prognostic workflow for building assets. Deep learning approaches require data volumes that individual building assets cannot provide (Hu et al., 2023; Hakimi, Liu and Abudayyeh, 2025). Ontological standards provide the data infrastructure but embed no quantitative prognostic capability (Balaji et al., 2018). Baseline methods generate the operational features but do not themselves estimate degradation. The present study addresses this gap by integrating all three elements (ontological grouping for data pooling, climate-matched baselines for feature generation, and a collectively trained machine-learning model for daily degradation estimation) into a single workflow that operates under the data scarcity conditions characteristic of the building sector.

From this body of literature, four scientific issues that motivate the present work can be made explicit. First, individual building assets accumulate insufficient degradation and failure data for standalone prognostic modelling, and the resulting class imbalance biases supervised models toward healthy operation (Hu and Cai, 2024; Hosamo et al., 2022). Second, the degradation rate of a building asset is non-stationary because it depends on climate, occupancy, and maintenance quality rather than on age alone, so models trained on static conditions lose predictive power when the operating context shifts (Hosseini Gourabpasi and Nik-Bakht, 2024; Kang, Chung and Hong, 2021). Third, ontological standards such as Brick provide the data infrastructure for cross-asset interoperability but embed no quantitative prognostic capability (Balaji et al., 2018). Fourth, deep sequence models such as LSTM require continuous multi-year recordings that are unavailable in newly instrumented buildings and tend to overfit when the underlying dataset is small (Hu et al., 2023; Hakimi, Liu and Abudayyeh, 2025). The workflow presented in Section 3 addresses these four issues through the integrated design introduced above.

3. METHOD

3.1. Architecture overview

Based on the data scarcity analysis in Section 2, the workflow adopts Random Forest as its core predictive model for its resistance to overfitting with small datasets, its natural handling of tabular daily features, and its built-in uncertainty quantification through inter-tree variance. This choice is consistent with the most recent comparative evidence on machine-learning algorithms for building predictive maintenance, where Random Forest is reported as the most widely adopted method, achieving accuracies of 89 to 92 percent and approximately 96 percent lower runtime than gradient-boosted alternatives (Hosamo et al., 2023; Abdelalim et al., 2025).

Rather than modelling degradation as a continuous time-series, the method estimates how much an asset degrades on each individual day from its operating conditions. Climate data serves as a proxy to generate usage baselines that forecast expected energy consumption, temperature differentials, and stress levels. The daily degradation estimates are accumulated over time: when the cumulative total reaches 100 percent, the number of remaining days is the predicted Remaining Useful Life. When real-time consumption data is available, the baseline is replaced by actual measurements; the workflow then operates in forecast mode (climate-based baselines) or monitoring mode (live data), both feeding the same predictive model.

The four-level architecture itself is, however, agnostic to the choice of regressor: any tabular machine-learning model can be substituted in Level 2 without altering Equations (1) to (5) or the structure of Levels 1, 3 and 4. The comparative evaluation of alternative regressor families is identified in Section 6 as a future-work priority.

This section describes the processing pipeline. The architecture is organised into four processing levels, illustrated in Figure 1. Level 1 (baseline calculation) ingests asset historical data (energy consumption, internal temperatures and external weather forecasts) and generates daily usage baselines using the X-of-Y climate-matching algorithm described in Section 3.2. If new real-time operational data is available, it replaces the baseline forecast for the corresponding days. Level 2 (prognostic engine) contains two sub-processes: a training module that constructs the collectively trained machine-learning model from pooled asset data, and a daily prediction module that applies the trained model to the baseline-derived features to estimate the asset-specific daily degradation rate and its associated uncertainty. Level 3 (verification) compares the accumulated sum of daily degradation rates (ASDR) against the asset's total useful life and checks whether the iteration limit has been reached; this is a simple convergence check

that determines whether the RUL projection has stabilised or requires further daily increments. Level 4 (human-in-the-loop) closes the feedback loop: when a maintenance intervention occurs, the asset's condition is reassessed, the maintenance report updates the database, and the training module retrains the model with the new information, progressively improving prediction accuracy over successive maintenance cycles.

The four-level architecture operates on top of a semantic layer that uses the Brick schema (Balaji et al., 2018) as the controlled vocabulary for asset classification. Each physical unit deployed in the building portfolio is mapped to a single Brick equipment class (for example, brick:Fan_Coil_Unit, brick:Air_Handling_Unit, brick:Pump) and is associated with the Brick relationships that locate it within the building topology (hasLocation, feeds, isPartOf). This semantic grounding has two functional consequences for the prognostic workflow. First, the training module groups assets for collective training strictly by Brick equipment class, ensuring that pooled samples come from functionally homogeneous units rather than from superficially similar but mechanically distinct equipment. Second, the resulting models are stored under their Brick class identifier, so a model trained on fan coil units in one building can be applied to fan coil units in another without manual reconfiguration. The ontology layer therefore provides both the grouping criterion that defines the collective (pooled) training dataset and the portability mechanism that allows trained models to transfer across the building portfolio. The classification step that assigns each asset its Brick class and an initial Asset Health Index is treated as a precondition of the present workflow; its details are reported in the companion paper on cold-start criticality assessment.

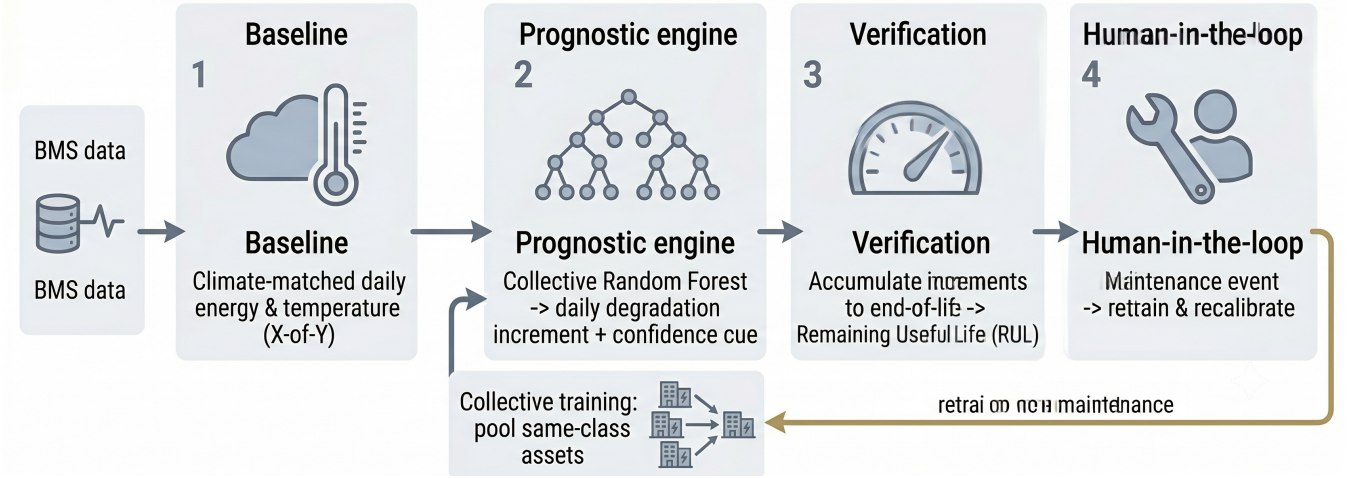

Figure 1. Four-level architecture of the collective learning workflow. Level 1 (baseline generation) implements the climate-matched X-of-Y forecasting of Equation (1). Level 2 (prognostic engine) contains the collective Random Forest of Equations (4) and (5) and produces the daily degradation increment of Equation (2). Level 3 (verification) accumulates daily increments according to Equation (3) and tests the convergence to the end-of-life condition that defines the RUL. Level 4 (human-in-the-loop) feeds maintenance events back into the training set and triggers periodic model retraining.

3.2. Baseline generation

The first processing stage generates daily usage baselines using historical operational data and outdoor temperature as a proxy. This is one of the workflow's distinguishing features: by forecasting the intensity and conditions under which an asset will operate, the tool produces degradation rates that reflect the actual demands placed on the equipment. A fan coil serving a heavily occupied building in a warm climate will degrade differently from one in a mild climate with low demand, even if both are identical in model and age.

The selected baseline algorithm is X of Y. For each day to be predicted, the method identifies the X days from the Y most recent historical days whose mean outdoor temperature is closest to the target day, and averages their energy consumption and temperature values. This approach was chosen for its simplicity, low computational cost, and effectiveness with the limited multi-year data available at the pilot site. Preliminary validation yielded a monthly CVRMSE of approximately 3 percent for temperature, within the range reported in the literature for this type of baseline. Energy consumption was more variable, at approximately 24 percent at monthly aggregation; this exceeds the literature range for the less intensively operated units, and Section 4 examines the cause.

$$\hat{B}(d) = \frac{1}{X} \sum_{i \in S_X(d)} V(d_i) \quad (1)$$

Where d is the target day for which the baseline is being predicted. V represents the recorded operational value (energy consumption or temperature) on a given historical day. T is the mean outdoor temperature. The algorithm searches the Y most recent historical days and selects the X days whose outdoor temperature is closest to the target day's temperature. The baseline prediction is the average of the operational values recorded on those X selected days. This approach assumes that days with similar outdoor temperatures will produce similar asset usage patterns.

A known limitation is that climate-based baselines cannot anticipate unprecedented disruptions such as a pandemic, where building occupancy drops to near zero. However, as new operational data is collected, the baselines are updated and progressively incorporate unusual patterns. The baseline output feeds directly into the degradation module as the expected operating conditions for each future day.

The data integration for baseline generation combines daily fan power sum, cumulative power consumed, mean air temperatures and external weather forecasts. By evaluating historical analogues, the X-of-Y algorithm calculates the expected daily usage parameters.

3.3. Collective model training

The training module produces the collective machine-learning model that the degradation module uses to estimate daily degradation. The training data is constructed as follows: for a given asset, the total degradation between two consecutive maintenance reports is known (from the AHI difference). The training module distributes this total degradation across the individual days in that period, weighting each day proportionally to its usage intensity. This creates a set of labelled samples where each day is characterised by its usage features (from the baseline module or from real data) and an associated degradation value. The collective machine-learning model is then trained on this dataset.

Each trained model is saved with a unique identifier linked to its asset class. This means that a model trained on fan coils from one building can be applied to fan coils in another, which is the mechanism through which the workflow implements collective learning: assets of the same class share their degradation history to compensate for the scarcity of data at the individual level. Every time a new maintenance report arrives, the training module retrains the model with the updated information, making the framework progressively more accurate.

This mechanism assumes that units of the same Brick class are prognostically interchangeable. Differences in manufacturing quality, build tolerance or unit cost between nominally identical units are not modelled explicitly, and are reflected only indirectly through the per-unit condition rating that anchors each asset's degradation label. The extent to which this assumption holds across manufacturers and procurement batches is a limitation examined in Section 5.

$$\delta(d) = \Delta\text{AHI} \times \frac{u(d)}{\sum_{i=t_1}^{t_2} u(i)} \quad (2)$$

where the total change in Asset Health Index between two consecutive maintenance reports is distributed across the days in that period. Each day receives a share of the total degradation proportional to its usage intensity u(d), which for fan coils corresponds to daily energy consumption. The denominator is the sum of all daily usage values in the period, ensuring that the individual daily values add up exactly to the total degradation. This produces a labelled dataset where each day is characterised by its operational features (usage, stress, fatigue) and an associated degradation target, which is then used to train the collective machine-learning model.

The inputs are the AHI values and the daily usage data. The output is a trained model file stored with a unique class identifier.

Internally, the machine-learning training module processes historical usage data alongside the calculated Asset Health Index, and aggregates predictions from multiple decision trees to form the final regressor model.

3.4. Daily degradation and RUL projection

The degradation module is the core of the predictive engine. It takes three inputs: the Asset Health Index (the starting condition), the daily usage baselines (the expected operating conditions), and a trained machine-learning model (the learned relationship between usage and degradation). For each day, the model receives the usage, stress, and fatigue features and outputs an estimated degradation increment. These increments are accumulated starting from the AHI: when the cumulative degradation reaches 100 percent, the asset has reached its end of life. The number of days between the current date and that point is the predicted RUL.

The Asset Health Index that enters the degradation module is, in the present formulation, a two-variable construct derived from the chronological age of the asset and its most recent physical condition rating. This minimal definition is a deliberate choice for the cold-start phase, when richer signals such as vibration spectra, acoustic baselines, or refrigerant-circuit pressure trends are typically unavailable at portfolio scale. The workflow, however, is agnostic to the internal definition of the AHI: any continuous index bounded in the [0, 100] interval can be substituted for the current formulation without altering Equations (2) and (3). A natural avenue for refinement is to extend the AHI to a multi-variable index that combines chronological age, condition rating, cumulative usage intensity, maintenance frequency, and, where available, condition-monitoring

signals such as fan-motor current trends or coil pressure-drop drift. This extension would propagate through the daily degradation module as a more informative starting condition for the cumulative trajectory, without requiring structural changes to the Random Forest model or to the collective training procedure.

When real-time operational data is available, the degradation module replaces the baseline forecast with actual measurements for each day that has already passed. This progressively refines the degradation curve: past days use real data, future days use the baseline prediction. As a result, the RUL estimate becomes more precise with each day of real operation.

$$D(t) = \mathrm{AHI}_0 + \sum_{i=t_0}^{t} \delta(i) \tag{3}$$

where *D(t)* is the cumulative degradation at day *t*, ranging from 0 (new) to 1 (end of life). The initial value is the Asset Health Index at the date of the last maintenance report. Each daily increment represents the degradation estimated by the machine-learning model for that day's operating conditions. For days already passed, actual operational data is used; for future days, the baseline prediction is used instead (shown with the hat notation in the RUL equation). The RUL at any point in time is the number of days until the accumulated degradation is projected to reach 1.0.
In the daily degradation estimation sequence, the workflow feeds the generated usage baselines and the Asset Health Index into the trained model, which then processes the operational features to compute the expected daily wear for each asset.

3.5. Predictive-model instantiation: Random Forest

The formulation reported below describes the Random Forest instantiation used in the pilot application of Section 4. The prediction for a given day is the mean of K individual tree outputs:

$$\hat{y} = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} f_k(\boldsymbol{x}) \tag{4}$$

A model-confidence indicator is obtained from the dispersion of predictions across the individual trees of the ensemble:

$$\mathrm{Var}(\hat{y}) = \frac{1}{\mathrm{K}-1} \sum_{\mathrm{k}=1}^{\mathrm{K}} (f_k(\boldsymbol{x}) - \hat{y})^2 \tag{5}$$

where K is the number of trees in the ensemble, x is the daily feature vector containing usage, stress and fatigue values and, in the augmented configuration, the static condition indicator, and f_k is the prediction function of tree k. The inter-tree dispersion in Equation (5) is interpreted here as a heuristic indicator of model confidence rather than as a calibrated statistical uncertainty: high dispersion flags input regions in which the trees disagree, typically because the collective training data covers those operating conditions sparsely. This quantity does not, by itself, constitute a calibrated predictive interval. Formal statistical guarantees would require complementary techniques such as the infinitesimal-jackknife variance estimator for Random Forests (Wager, Hastie and Efron, 2014) or conformal prediction wrappers (Shafer and Vovk, 2008). Both are compatible with the architecture proposed here and are identified in Section 6 as a priority for future calibration work. In the present workflow, the dispersion in Equation (5) is reported to facility managers alongside each RUL projection as a relative reliability cue, useful to prioritise additional monitoring on assets whose operating conditions fall outside the densely sampled regions of the training set.

4. PRELIMINARY PILOT APPLICATION

This section presents a preliminary application of the workflow to operational data from a pilot commercial building. Building-management-system streams at five-minute resolution are available for eleven fan coil units of identical type (Ability MAT270) installed in 2021, covering January 2025 to February 2026. One of the eleven units was excluded from the prognostic analysis because more than 85 percent of its days exhibited zero consumption, an insufficient number of non-zero observations to calibrate the model; the analysis was therefore conducted on the remaining ten units. The ground-truth AHI for each unit was obtained from the S2 health-index module of the workflow's reference implementation, which combines a documented condition rating (rated Satisfactory, On Notice or Good across the ten units) with asset age, criticality and importance attributes to yield a per-unit AHI; for the ten retained units this produced cumulative degradation values ranging from 13 to 40 points on the 0 to 100 scale. The application is presented as preliminary: the operational data post-dates the inspection, and the workflow has not yet been closed against a second maintenance event.

The Level-1 baseline of Section 3.2 was applied to predict daily energy and indoor temperature on the last 40 percent of each unit's observation window under an expanding-window protocol. Indoor temperature predictions met the ASHRAE G14 monthly CVRMSE threshold for every unit in the sample (median approximately 3 percent). Daily energy CVRMSE was more variable across units (median 81 percent daily and 24 percent at monthly aggregation under the vanilla variant), with the highest values concentrated in units of low operational intensity where the small CVRMSE divisor amplifies absolute errors. A sensitivity analysis revealed that the efficiency filter of the baseline introduces a systematic low-side bias of approximately 29 percentage points in NMBE, which a mean-of-top-Y variant without the filter removes. The implications of this finding are discussed in Section 5.

The Level-2 collective Random Forest was evaluated under leave-one-out cross-validation across the ten retained units, each unit held out in turn while the model trained on the remaining nine, which removes the dependence on a single arbitrary train and validation partition. Two feature configurations were compared, both using 200 trees of maximum depth 15. The operational-only configuration, with the eight usage, stress and fatigue features, did not generalise the per-unit degradation level: the mean absolute percentage error across the held-out units was 33 percent (median 26 percent), the coefficient of variation of the root-mean-square error (CVRMSE) was 42 percent and the coefficient of determination was negative (R^2 = -0.48), no better than predicting the population mean (Table 2). Augmenting the feature set with the static condition rating that anchors the label, giving nine features, improved generalisation substantially: the mean absolute percentage error fell to 19 percent (median 13 percent), the mean absolute error to 4.8 cumulative ΔAHI points, CVRMSE to 20 percent and R^2 to 0.67 (Table 2). Under both configurations the largest residuals fall on the units whose true ΔAHI departs furthest from the population mean (Table 1), confirming that operational signals alone do not encode the degradation level and that fusing a static condition indicator with the operational dynamics is required to recover it.

A physics-only baseline that ignores operational data was computed for benchmark. Across the ten units it yielded cumulative ΔAHI values in the range 6.5 to 19 (mean 13.2), with a mean absolute percentage error of approximately 52 percent against the per-unit ground truth. The baseline under-predicts uniformly because it ignores operational intensity; the collective RF, by contrast, recovers the per-unit degradation level only when the static condition indicator is included alongside the operational features (Section 4, Table 1).

FCU	Actual ΔAHI	Op-only ΔAHI	error (%)	Op+rating ΔAHI	error (%)
01	40.03	25.49	36.3	35.87	10.4
02	28.03	19.93	28.9	15.93	43.2
04	28.03	30.21	7.8	24.46	12.7
05	40.03	21.65	45.9	36.20	9.6
06	16.03	17.88	11.6	8.97	44.0
07	13.32	15.19	14.0	10.14	23.9
08	28.03	24.61	12.2	25.68	8.4
09	28.03	21.76	22.4	23.55	16.0
10	40.03	18.42	54.0	34.82	13.0
11	16.03	31.65	97.4	14.11	12.0

Table 1. Per-unit leave-one-out predictions for both feature configurations (N=10). Actual ΔAHI is the per-unit ground truth from the reference implementation's S2 health-index module; error is the absolute relative error of each prediction.

Metric	Op-only	Op+rating
MAPE (mean ± SD)	33.1% ± 27.5%	19.3% ± 13.5%
MAPE (median)	25.6%	12.9%
MAE (ΔAHI points)	9.38	4.79
RMSE (ΔAHI points)	11.78	5.55
CVRMSE	42.4%	20.0%
R^2 (across units)	-0.48	0.67

Table 2. Aggregate leave-one-out validation metrics for the two feature configurations (N=10). MAPE is the mean absolute percentage error, CVRMSE the coefficient of variation of the RMSE, and R^2 the coefficient of determination across units.

Several limitations qualify these preliminary results. With only ten units in a single class, the cross-validated per-unit dispersion cannot be interpreted as a calibrated generalisation estimate. The operational data window post-dates the inspection, the sample comprises a single asset class at a single building, and the workflow has not been closed against a second maintenance event. These limitations motivate the future-work priorities listed in Section 6.

A further methodological caveat applies. The Random Forest is trained on operational features (daily energy use, indoor temperature, climate variables and derived ratios), whereas the ground-truth ΔAHI labels are produced by the S2 health-index module from static asset attributes (rating, age, criticality and importance) that are not exposed to the model. With the static attributes withheld, the model cannot recover the label structure from operational patterns alone and regresses towards the population mean on unseen assets, which inflates the relative error for units whose true ΔAHI departs from that mean. Exposing the static condition rating to the model as an additional feature reduced this effect substantially (Section 4); enlarging the pool to additional asset classes and buildings, and replacing the static index with a sensor-derived health indicator, are expected to close the remaining gap.

5. DISCUSSION

The collective training strategy addresses a structural limitation of building-scale prognostics: individual assets rarely accumulate sufficient degradation data for standalone model training, yet populations of the same ontological class collectively span a wider range of operating conditions than any single unit. By pooling daily features across multiple assets, the workflow generates a training dataset whose diversity compensates for the brevity of each individual history. This strategy is viable when assets share similar degradation mechanisms, which the ontological grouping ensures. However, its effectiveness depends on the degree of operational homogeneity within a class: assets of the same type serving fundamentally different thermal zones or occupancy patterns may introduce noise rather than signal into the collective pool.

The climate-matched baseline is the critical link between raw operational data and the prognostic model. Its accuracy directly determines the quality of the daily feature set and, consequently, the reliability of the degradation estimate. The dual-mode architecture (forecast for future days, monitoring for past days) provides a natural mechanism for progressive refinement: as real measurements replace baseline predictions, the cumulative degradation trajectory converges toward the true value. This self-correcting behaviour is particularly important during the early deployment phase, when the baseline has been trained on limited historical data.

Some limitations merit discussion. The proportional distribution of total degradation across days (weighted by usage intensity) assumes that degradation scales linearly with energy consumption. In practice, threshold effects (e.g. thermal cycling, condensation) may cause disproportionate damage under specific operating conditions.

A further limitation concerns the interpretation of the dispersion measure of Equation (5). As discussed in Section 3.5, inter-tree dispersion is a useful heuristic to flag regions of the feature space that the collective training set has not adequately covered, but it does not constitute a calibrated predictive interval and should not be reported as such to facility managers. The framework therefore presents this quantity as a relative confidence cue, with the explicit caveat that the development of calibrated intervals via infinitesimal-jackknife or conformal methods is required before the prognostic outputs can support contractual or safety-critical maintenance decisions.

A further limitation concerns the transferability of a pooled model across nominally identical units. Collective training treats all assets of the same Brick class as prognostically interchangeable, yet differential manufacturing quality, build tolerance and unit cost can produce genuinely different degradation behaviour under identical operating conditions. The present workflow captures this only indirectly, through the static per-unit condition rating that anchors each label, and not as an explicit unit-level effect; where build quality varies systematically within a class, the pooled model may transfer less faithfully than the ontological grouping alone would suggest. Quantifying this effect requires condition histories across units of known provenance, which the present single-class, single-building sample cannot provide.

The preliminary application surfaced two findings of methodological interest. First, the baseline's efficiency filter, designed to provide a counterfactual estimate of well-operated consumption, introduced a systematic low-side bias when used as a forecast estimator of observed consumption. A mean-of-top-Y variant without the filter removed the bias and brought part of the sample within the ASHRAE G14 monthly threshold. The two configurations therefore address different questions: the filtered variant estimates expected efficient operation as a target, whereas the unfiltered variant estimates expected operation as a forecast; Section 6 lists this distinction as a future refinement of the implementation. Second, under leave-one-out cross-validation the operational-only model did not generalise the per-unit degradation level (R^2 = -0.48). This is consistent with the static asset attributes that generate the label being absent from an operational-only feature set. Re-introducing the condition rating as a model feature raised the cross-validated R^2 to 0.67 and roughly halved the mean error, which confirms the diagnosis; closing the remaining gap motivates a sensor-derived health indicator, listed in Section 6. Fuller validation against per-unit follow-up inspections is required to disambiguate prognostic capture from training-set extrapolation.

6. CONCLUSION

This paper presented a prognostic workflow for building-asset remaining useful life estimation under the data-scarcity conditions typical of the sector. It combines the three elements of Section 2.5: climate-matched usage baselines that supply daily features without multi-year recordings, a collectively trained model that pools data across ontologically homogeneous assets, and a daily calculation cycle with human-in-the-loop recalibration. RUL projections carry an uncalibrated dispersion cue from the variance across trees, with calibrated intervals left as future work.

A preliminary application to fan coil units at a pilot building recovered the per-unit labels with a cross-validated R^2 of 0.67 (CVRMSE 20 percent) once the static condition indicator was fused with the operational features, whereas an operational-only model did not generalise (R^2 below zero).

Six priorities remain. First, fuller validation with more held-out units, follow-up condition ratings and multi-season horizons, expected to reduce the regression-to-the-mean seen under leave-one-out cross-validation. Second, a sensitivity analysis of the X-of-Y baseline, AHI weighting and efficiency filter to set robust deployment defaults. Third, extension to further asset classes (air handling units, pumps, chillers) and building typologies to map where ontological grouping helps. Fourth, an end-to-end test of the Level-4 human-in-the-loop cycle against a second maintenance event. Fifth, benchmarking alternative tabular regressors (gradient-boosted trees, neural-network ensembles, transformer-based models) against the Random Forest to match a model family to each equipment class. Sixth, modelling unit-level heterogeneity explicitly, for example by adding manufacturer, model variant or production batch as categorical features, or by adopting a hierarchical (mixed-effects) formulation with per-unit random effects, so that systematic differences in build quality within a Brick class can be separated from operational degradation.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of the WILSON project (May 2024 - April 2028), funded by the European Union under Grant Agreement ID: 101147267, within the Horizon Europe framework (Call: HORIZONCL5-2023-D4-02-01).

REFERENCES

Abdelalim, A.M., Essawy, A., Sherif, A., Salem, M., Al-Adwani, M., & Abdullah, M.S. (2025). Optimizing facilities management through artificial intelligence and digital twin technology in mega-facilities. Sustainability, 17(5), 1826.

ASHRAE (2014). Guideline 14-2014: measurement of energy, demand, and water savings. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Acer, Y., Berges, M., Culler, D., Gupta, R., Kjaergaard, M.B., Srivastava, M., & Whitehouse, K. (2018). Brick: metadata schema for portable smart building applications. Applied Energy, 226, 1273-1292.

Asare, K.A.B., Liu, R., & Anumba, C.J. (2025). Generative AI for predictive maintenance in buildings. In Proceedings of SASBE 2024 (pp. 928-935). Springer.

Guzmán-Torres, J.A., Domínguez-Mota, F.J., Alonso-Guzmán, E.M., Martínez-Molina, W., & Tinoco-Guerrero, G. (2025). A digital twin approach based method for classification of salt damage in building evaluation. Mathematics and Computers in Simulation, 233, 433-447.

Hakimi, O., Liu, H., & Abudayyeh, O. (2025). Deep learning-driven multi-level data fusion for predictive maintenance of concrete bridge decks. Automation in Construction, 175, 106180.

Hosamo, H.H., Svennevig, P.R., Svidt, K., Han, D., & Nielsen, H.K. (2022). A digital twin predictive maintenance framework of AHUs based on automatic fault detection and diagnostics. Energy and Buildings, 261, 111988.

Hosamo, H.H., Nielsen, H.K., Kraniotis, D., Svennevig, P.R. and Svidt, K., 2023. Digital Twin framework for automated fault source detection and prediction for comfort performance evaluation of existing non-residential Norwegian buildings. *Energy and Buildings*, *281*, p.112732.

Hosseini Gourabpasi, A., & Nik-Bakht, M. (2024). BIM-based automated fault detection and diagnostics of HVAC systems in commercial buildings. Journal of Building Engineering, 87, 109022.

Hu, W., & Cai, Y. (2024). A semi-supervised method for digital twin-enabled predictive maintenance. Neural Computing and Applications.

Hu, W., Wang, X., Tan, K. and Cai, Y., 2023. Digital twin-enhanced predictive maintenance for indoor climate: A parallel LSTM-autoencoder failure prediction approach. *Energy and Buildings*, *301*, p.113738..

Hyndman, R.J., & Koehler, A.B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

Kang, J.S., Chung, K., & Hong, E.J. (2021). Multimedia knowledge-based bridge health monitoring using digital twin. Multimedia Tools and Applications, 80, 34609-34624.

Ma, X., Shi, Z., Guo, F., Chen, Z., & Wei, J. (2023). Digital twin model for chiller fault diagnosis based on SSAE and transfer learning. Building and Environment, 243, 110718.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371-421.

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for Random Forests: the jackknife and the infinitesimal jackknife. Journal of Machine Learning Research, 15, 1625-1651.

.