From Interoperability to Explainability: Semantic Data Models for XAI in Industrial Prognostics and Health Management

Anand Todkar[1], Mrinmoy Sarkar[1], Jitendra Solanki[1], Ziran Min[1]

[1] *Siemens Foundational Technology, Princeton, NJ, USA*
anand.todkar@siemens.com
mrinmoy.sarkar@siemens.com
jitendra.solanki@siemens.com
ziran.min@siemens.com

ABSTRACT

Semantic data models play a key role in enabling interoperability in industrial plants, and their importance has grown further with the rapid adoption of industrial artificial intelligence (AI). As AI applications become increasingly prevalent, most of them rely on data-driven approaches and exhibit a predominantly black-box nature, making their output difficult to interpret and explain. To address this limitation, this work investigates the integration of semantic data models with AI applications to enhance the interpretability of their results.

In this study, the OPC UA ISA-95 companion specification is utilized as the semantic data model, while a random forest–based AI model is applied for predictive maintenance in a manufacturing shop floor context. To support explainable AI (XAI), the results are evaluated in five dimensions: faithfulness, stability, sparsity, redundancy, and calibration. Furthermore, a comparative analysis is conducted under three scenarios: (i) using a flat tabular data set with feature engineering, (ii) incorporating the semantic model, and (iii) augmenting the semantic model with the flat tabular data set. Preliminary findings indicate that the integration of semantic data models enhances the explainability of black-box AI models.

1. INTRODUCTION

Predictive maintenance represents a high impact application of machine learning (ML) in industrial systems. Machine Learning based methods have been widely investigated for detecting and predicting degradation mechanisms such as bearing wear, pump cavitation, and motor failure (Lei et al., 2018; Zhang, Yang, & Wang, 2019; Serradilla, Zugasti, Rodriguez, & Zurutuza, 2022). Because maintenance decisions can affect operational safety, equipment availability, and production costs, accurate predictions alone are often insufficient. Operators, engineers, and regulators increasingly require interpretable evidence indicating which sensor measurements contributed to a fault prediction, which physical subsystems were implicated, and how confidently the explanation can be interpreted (Vollert, Atzmueller, & Theissler, 2021; Nor, Pedapati, Muhammad, & Leiva, 2021). These requirements motivate the use of eXplainable Artificial Intelligence (XAI), which aims to make the behavior and outputs of ML models inspectable by human stakeholders.

A common industrial XAI workflow represents sensor measurements as a flat set of input features, trains a predictive model on this representation, and subsequently applies a post-hoc explanation method such as SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017; Lundberg et al., 2020). The resulting explanations assign importance values to individual input features. Although such attributions identify influential sensor variables, they generally do not represent the structural relationships among the sensors, components, and equipment to which those variables belong. Consequently, the explanations remain disconnected from the physical and organizational structure of the industrial asset.

This limitation is notable because industrial interoperability standards already provide machine-readable asset models. ISA-95, standardized through IEC 62264, defines a hierarchical organization of industrial assets across the levels of *enterprise*, *site*, *area*, *line*, and *equipment* (International Electrotechnical Commission, 2013). Similarly, OPC UA (Open Platform Communications Unified Architecture) provides a standardized framework for machine-to-machine communication and information modeling in industrial environments (International Electrotechnical Commission, 2020). Despite the availability of this semantic information, conventional ML pipelines typically discard the asset hierarchy during preprocessing and retain only a flat matrix of sensor values.

This study investigates whether preserving and incorporating the semantic structure of an industrial asset model improves the quality of post-hoc ML explanations. The proposed approach is motivated by the Siemens Industrial Information Hub (IIH), a SIMATIC Industrial Edge application that exposes industrial data through semantically structured OPC UA information models. In particular, the IIH Semantics component publishes a typed OPC UA address space, while the OPC UA *Aggregation Server* consolidates plant-wide information represented using OPC UA companion specifications (Siemens AG, 2025, 2024). Companion specifications extend the OPC UA framework with standardized, domain-specific information models. Examples include ISA-95, PackML for packaging machinery, and OPC UA Robotics.

In contrast to a flat comma-separated values (CSV) representation, an IIH-based OPC UA address space can associate each process variable with its equipment path, engineering unit, physical-asset hierarchy, and component group. The semantic features considered in this work are not manually engineered from external domain knowledge. Instead, they are obtained as algebraic summaries of the ISA-95 component groups already materialized by the OPC UA companion specification. For an IIH-equipped industrial system, these component assignments, equipment paths, and engineering units are therefore available as a by-product of the existing interoperability infrastructure rather than as the result of an additional manual modeling activity.

The experimental study uses the ISA-95 companion specification as a concrete realization of the proposed methodology. However, the method is not specific to ISA-95 and can be applied to other companion specifications accessible through the same OPC UA Aggregation Server. To isolate the effect of the input representation, the predictive model and training procedure are held fixed, while only the feature encoding is varied. A conventional flat sensor representation is compared with an ISA-95-aware semantic representation using a one-to-one correspondence between the underlying sensor attributes.

Explanation quality is evaluated along five quantitative dimensions: faithfulness, stability, sparsity, redundancy, and calibration. These dimensions are implemented using metrics supported by the Quantus evaluation toolkit and are motivated by established studies on the evaluation of XAI methods (Hedström et al., 2023; Bhatt, Weller, & Moura, 2020; Samek, Binder, Montavon, Lapuschkin, & Müller, 2017; Alvarez-Melis & Jaakkola, 2018). This evaluation enables a systematic assessment of whether industrial semantic structure improves not only the interpretability of explanations to human operators, but also their measurable reliability and information quality.

The principal contributions of this work are as follows:

1. We define a one-to-one feature-encoding methodology for comparing conventional flat sensor attributes with an ISA-95-aware semantic OPC UA representation while holding the predictor and underlying measurements fixed.
2. We propose a five-dimensional quantitative framework for evaluating industrial XAI explanations in terms of faithfulness, stability, sparsity, redundancy, and calibration.
3. We provide an empirical assessment of whether semantic information already available through OPC UA companion specifications can improve XAI quality without requiring additional manual feature engineering or asset modeling.

2. RELATED WORK

Industrial XAI lies at the intersection of semantic data representation, post-hoc explanation, and quantitative evaluation. Accordingly, this section reviews how industrial asset semantics are represented through ISA-95 and OPC UA, how XAI methods are applied to industrial sensor data, and how explanation quality is evaluated. This provides the basis for examining whether semantic feature encoding can improve explanations for industrial fault-prognostics models.

2.1. Semantic Interoperability in Industrial Systems

Semantic interoperability requires both a common representation of industrial assets and a mechanism for exchanging that representation between heterogeneous systems. ISA-95 provides the organizational foundation by defining a hierarchy of industrial assets across the *enterprise*, *site*, *area*, *line*, and *equipment* levels. In addition, OPC UA provides the corresponding communication and information-modeling framework. The companion specification ISA-95 connects these standards by translating the ISA-95 hierarchy into standardized OPC UA types, including `EquipmentType` and `PhysicalAssetType`. As a result, industrial measurements can be represented together with their asset hierarchy, component membership, and engineering context rather than as isolated sensor values.

Siemens Industrial Information Hub (IIH) operationalizes this semantic architecture on SIMATIC Industrial Edge. Its semantics component models industrial assets, while its OPC UA Aggregation Server consolidates data represented through companion specifications such as ISA-95, PackML, and OPC UA Robotics. IIH therefore provides off-the-shelve platform for transformingdistributed sensor measurements to a unified, semantically structured address space that can support downstream analytics.

Previous work has primarily studied this infrastructure from an interoperability perspective, including OPC UA communication, Industrial Internet of Things integration, IT/OT inte-

gration and automated asset discovery (Profanter, Tekat, Dorofeev, Rickert, & Knoll, 2019; Schleipen, Gilani, Bischoff, & Pfrommer, 2016; Mazzetti et al., 2020; Todkar, Sarkar, & Solanki, 2025; Todkar, Sarkar, Solanki, & Tylka, 2025). These studies establish how industrial semantics can be created and exchanged, but do not examine how that structure affects ML models that consume the resulting data. In particular, whether ISA-95-aware representations improve the faithfulness and reliability of post-hoc explanations remains underexplored.

2.2. XAI for Industrial Data

Industrial prognostics increasingly uses post-hoc XAI to identify the sensor measurements contributing to predicted equipment faults. SHAP is widely adopted for this purpose (Lundberg & Lee, 2017; Lundberg et al., 2020; Serradilla et al., 2022; Nor et al., 2021; Vollert et al., 2021), while LIME and Anchors provide surrogate- and rule-based alternatives (Ribeiro, Singh, & Guestrin, 2016, 2018).

Most industrial XAI models, however, operate on flat sensor tables that omit equipment hierarchies, component relationships, and engineering metadata. Their explanations can identify influential sensors but cannot directly relate them to the corresponding physical subsystems. This study examines whether retaining ISA-95 and OPC UA semantics during feature encoding produces more informative and quantitatively reliable explanations.

2.3. Quantitative XAI Evaluation

Although SHAP is widely used in industrial applications, attribution plots alone provide limited evidence of explanation quality. Quantitative evaluation methods have therefore been developed to test whether explanations reflect model behavior, remain stable under small perturbations, and concentrate importance on relevant features (Samek et al., 2017; Alvarez-Melis & Jaakkola, 2018; Yeh, Hsieh, Suggala, Inouye, & Ravikumar, 2019; Bhatt et al., 2020). Deletion and insertion tests, sanity checks, and adversarial robustness analyses provide complementary measures of these properties (Petsiuk, Das, & Saenko, 2018; Hooker, Erhan, Kindermans, & Kim, 2019; Adebayo et al., 2018; Slack, Hilgard, Jia, Singh, & Lakkaraju, 2020). Many such metrics are available through the Quantus toolkit (Hedström et al., 2023). However, quantitative XAI evaluation remains uncommon in industrial prognostics, where explanations are typically presented as SHAP plots supported by qualitative interpretation. This study applies a multi-dimensional evaluation to determine whether semantic feature encoding improves explanation faithfulness, stability, sparsity, redundancy, and calibration.

Our literature review indicates that prior work has examined semantic interoperability, post-hoc XAI, and quantitative explanation evaluation largely as separate topics. This study connects these areas by evaluating whether semantic asset information available through ISA-95-aware OPC UA models improves explanations for industrial fault classification.

The study pursues three aims:

1. Determine the effect of feature encoding on explanation quality while holding the observations, predictor, and explanation method fixed;
2. Assess whether ISA-95 semantics improve downstream ML explanations; and
3. Quantify differences in faithfulness, stability, sparsity, redundancy, and calibration.

We compare flat, semantic-only, and semantic-augmented encodings of identical observations using the same random forest configuration and stratified train–test split. An `asyncua`-based ISA-95 test bench materializes the semantic representations and emulates the relevant IIH data path. Explanation quality is evaluated across the five dimensions listed above, while differences in predictive performance are assessed using paired bootstrap tests.

To our knowledge, this is the first controlled study to hold the predictor fixed while varying the representation of the characteristics with respect to the semantics of ISA-95 and OPC UA and quantitatively evaluating the resulting explanations.

3. METHODOLOGY

Figure 1 shows the end-to-end pipeline; the subsections below detail each phase.

3.1. Dataset Preparation

We use the publicly available *Pump Sensor Data* dataset (Elgendy, 2018), a 220 320-row, 1-minute-cadence multivariate time-series capture from an instrumented motor-pump unit (52 sensors and three operational labels: NORMAL 93.4%, RECOVERING 6.6%, BROKEN 0.003%). The dataset ships as a flat CSV with no asset hierarchy attached – it is representative of the historian-style export that motivates this study – and we re-format it under an ISA-95 information model (Section 3.B) to expose it through OPC UA. Missing values are handled with forward/back-fill imputation. Following common practice for severely imbalanced multiclass prognostics (Lei et al., 2018; Zhang et al., 2019), the target is collapsed to binary abnormal-vs-normal (6.57% positive). The fully-null column `bearing_temperature_4` is dropped at modelling time but kept in the OPC UA address space.

3.2. ISA-95 Information Model

The unit `Motor-Pump-Unit-01` is modelled under ENTERPRISE → SITE → AREA → LINE → EQUIPMENT (5 levels), with six sub-equipment ob-

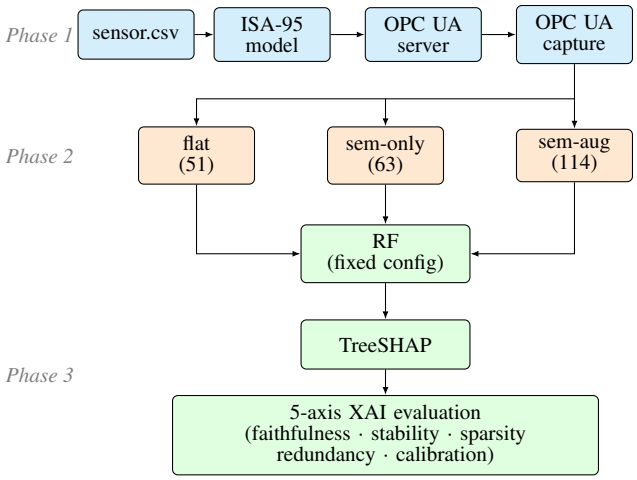

Figure 1. Experimental pipeline. **Phase 1**: the raw CSV is lifted into an ISA-95 information model and streamed through an OPC UA server (IIH analogue) to capture a roundtrip trace (`phase1_summary.json`); residual numeric drift is bounded and reported in Eq. (1). **Phase 2**: three feature representations are derived from the captured semantic metadata. **Phase 3**: a single fixed-configuration Random Forest and TreeSHAP are applied once per representation; the five-axis XAI metric suite grades the resulting explanations.

jects (`Motor`, `Shaft`, `Pump`, `LubricationSystem`, `ProcessLoad`, `GlobalMonitoring`). The 52 sensors are exposed as OPC UA `BaseDataVariable` nodes; each carries an `EngineeringUnits` string property, a `ComponentGroup` property (one of 12 ISA-95 component groups), and an `ISA95Path` property. Equipment nodes carry the `EquipmentLevel` property required by OPC 10030 (OPC Foundation, 2022). The IIH-analogue bench is intentionally minimal: it uses generic OPC UA objects and string-valued metadata properties rather than the typed `EquipmentType`/`PhysicalAssetType` nodes and `EUInformation` structures that a fully conformant companion-server (such as IIH) would emit. The metadata required to instantiate the typed nodes (paths, units, groups) is preserved end-to-end and made available to the consumer; conformance to the ISA-95 companion specification is therefore a packaging step rather than an information change. The model lives in `poc/asset_model.py`.

3.3. OPC UA Test Bench (IIH Analogue)

An `asyncua` 2.0 server (`poc/opcua_server.py`) hosts the address space and replays the dataset row-by-row. A separate client process (`poc/opcua_capture.py`) subscribes to all 52 variables plus a `row_index` trigger that the server bumps after each row's writes; this gives the client an atomic snapshotting cue. Captured rows are written in long form with full ISA-95 decoration (`enterprise`, `site`, `area`, `line`, `equipment`, `sub_equipment`, `component_group`, `node_id`, `isa95_path`, `engineering_unit`).

We pivot the captured long-form table back to wide form by `(row_index, browse_name)`, align with the source CSV's first 200 rows, and compute the per-column max absolute difference as in Eq. (1).

$$\delta_c = \max_{i \in 1..200} \left| \hat{x}_{i,c} - x_{i,c} \right|, \quad \bar{\delta} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \delta_c \tag{1}$$

Across the 52 captured columns we observe $\max_c \delta_c = 79.17$ and $\bar{\delta} = 8.55$ (`Data/phase1_summary.json`). The residual is non-zero – the bench at this stage demonstrates the IIH-portable *transport path* (asset-model lift, OPC UA hosting, client-side capture, long-to-wide reconstruction with full ISA-95 decoration) rather than bit-for-bit fidelity; isolating the remaining drift (likely a row-alignment or floating-point-encoding effect) is left to follow-up work and does not bear on the encoding-vs-explanation comparison studied in Sections 4–5, which uses the source CSV directly.

3.4. Three Feature Representations

- *flat*: 51 raw sensor columns (`bearing_temperature_4` excluded).
- *semantic-only*: 63 derived features = 12 components × 5 stats (mean / std / max / min / range) + 3 equipment-level aggregates.
- *semantic-augmented*: 114 features (the union).

Construction is in `poc/build_semantic_dataset.py`; the per-component aggregate is a leakage-free algebraic summary of values already in the same row.

3.5. Predictor

A single fixed Random Forest configuration is used in all three runs: `n_estimators=120`, `max_depth=12`, `class_weight=balanced_subsample`, `random_state=42`. The same stratified 80/20 split (`random_state = 42`) is reused, so the test indices are byte-identical across the encodings; this enables for paired-bootstrap analysis.

3.6. XAI Evaluation Metrics

Five axes are computed on a 200-row stratified evaluation set (100 positive, 100 negative). Per-feature attributions are produced by TreeSHAP – the exact, polynomial-time variant of SHAP for tree-ensemble models (Lundberg et al., 2020) – with a 100-row reference *background* set used to define the baseline expectation.

(a) Faithfulness (Petsiuk et al., 2018; Samek et al., 2017): order features descending by $|\mathrm{SHAP}|$; sweep k from 0 to n_{feat} in

Table 1. Random Forest, full-data stratified 80/20 split. 95% CIs from 500 bootstrap resamples.

Encoding	#feat.	F1 [95% CI]	PR-AUC
flat	51	.9964 [.9947,.9977]	1.000
semantic-only	63	.9981 [.9968,.9991]	1.000
semantic-augmented	114	**.9988** [.9978,.9997]	1.000

Table 2. Paired-bootstrap deltas (same test indices, 500 resamples, two-sided p).

Comparison	ΔF1 [95% CI]	p
sem-only vs. flat	$+.00174$ $[+.0002, +.0033]$	0.036
sem-aug vs. flat	$+.00242$ $[+.0012, +.0037]$	< 0.001

10 steps; record $\overline{P(\hat{y}=1)}$ when the top-k features are zeroed (DELETION) or restored from a zero baseline (INSERTION); report the trapezoidal AUC. **(b) Stability** (Alvarez-Melis & Jaakkola, 2018): empirical Lipschitz constant under additive Gaussian noise of 1% per-feature standard deviation. **(c) Sparsity at two resolutions** (Bhatt et al., 2020): number of features and number of ISA-95 components needed to capture 90% of |SHAP| mass per sample. **(d) Redundancy stress**: three sensors duplicated with 5%-noise copies; the Random Forest is retrained; the fraction of |SHAP| that migrates from the original to the duplicate is reported. **(e) Calibration** (Guo, Pleiss, Sun, & Weinberger, 2017): 15-bin expected calibration error (ECE), the mean discrepancy between predicted probability and empirical accuracy across bins.

4. RESULTS

4.1. Predictive Performance (H1)

Table 1 reports point metrics with bootstrap 95% confidence intervals (CIs). Paired-bootstrap deltas (Table 2) confirm that semantic encodings do not degrade – and slightly improve – the F1 score. The precision-recall area under the curve (PR-AUC) is saturated.

H1 is supported.

4.2. Faithfulness (H2)

Deletion AUC drops sharply (Table 3). Relative reduction:

$$\frac{0.4506 - 0.3482}{0.4506} = 22.7\% \quad \text{(sem-only vs flat)},$$
$$\frac{0.4506 - 0.3225}{0.4506} = 28.4\% \quad \text{(sem-aug vs flat)}.$$

The faithfulness score INS$-$DEL improves from -0.173 to -0.066, a $+62\%$ relative improvement. **H2 is supported.**

Figure 2 plots the deletion curves for all three encodings; the visual separation makes clear why the semantic-augmented model has the lowest deletion AUC.

Table 3. Faithfulness (200-row stratified evaluation set).

Encoding	DEL $\downarrow$	INS $\uparrow$	INS$-$DEL $\uparrow$
flat	0.4506	0.2772	-0.1734
semantic-only	0.3482	0.2492	-0.0990
semantic-augmented	**0.3225**	0.2561	$\mathbf{-0.0664}$

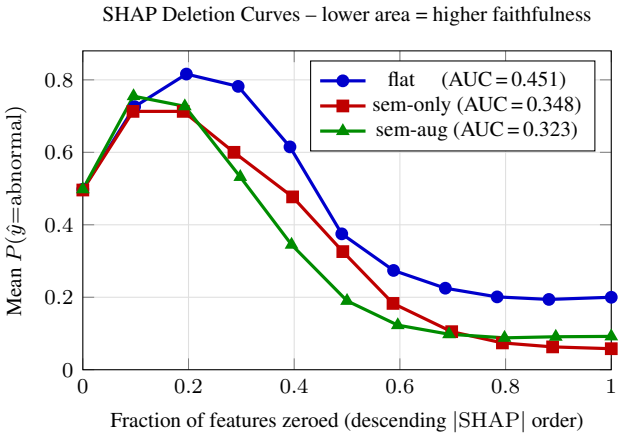

Figure 2. SHAP deletion curves (200-row stratified evaluation set, 100 positive / 100 negative). Features are zeroed progressively in descending |SHAP| order; a lower area under the curve (AUC) indicates that the top-ranked features are genuinely decision-critical, i.e. the explanation is more faithful. The semantic-augmented encoding achieves a 28.4% lower deletion AUC than flat (Table 3).

4.3. Stability

Mean Lipschitz under 1% Gaussian noise: $L_{\text{flat}} = 0.00243$, $L_{\text{sem-only}} = 0.00164$, $L_{\text{sem-aug}} = 0.00116$. Relative improvement vs. flat: $(0.00243 - 0.00116)/0.00243 = 52.3\%$. All encodings are very stable in absolute terms.

4.4. Sparsity at Two Resolutions (H3)

The result splits across the two semantic encodings. The semantic-only encoding requires $6.39/5.71 = 1.119\times$ ($+11.9\%$) more components to reach 90% mass than flat: dropping the raw sensors removes attribution paths the model otherwise exploits, and component-level conciseness suffers as a result. The semantic-augmented encoding, by contrast, has $114/51 = 2.24\times$ as many features as flat yet requires only $5.80/5.71 = 1.016\times$ as many components for the same 90% mass: a $+1.6\%$ change at the component level versus a $+124\%$ feature-space increase (Table 4). **H3 is supported for the semantic-augmented encoding**, which is also the encoding that delivers the headline faithfulness, stability, and calibration gains; the semantic-only encoding does not preserve component-level conciseness in this dataset.

Table 4. Number of features and ISA-95 components needed to capture 90% of |SHAP| mass.

Encoding	feats. for 90%	comps. for 90%
flat	15.5	5.71
semantic-only	22.0	6.39
semantic-augmented	31.6	5.80

Table 5. Cross-axis summary; arrows indicate desired direction. Relative gain compares the best semantic encoding to the flat baseline.

Axis	flat	best-sem.	rel. gain
F1 (↑)	0.9964	0.9988	$+0.24\%$
Deletion AUC (↓)	0.4506	0.3225	$\mathbf{-28.4\%}$
INS−DEL (↑)	−0.173	−0.066	$\mathbf{+61.7\%}$
Lipschitz (↓)	0.0024	0.0012	$\mathbf{-52.3\%}$
Comp. for 90% (↓)	5.71	5.80	$+1.6\%$
Redundancy stolen (↓)	0.420	0.423	$+0.7\%$
ECE_{15} (↓)	0.0016	0.0013	-18.8%

4.5. Redundancy Stress

Stolen-fraction under 5%-noise duplicates of three representative sensors (`motor_bearing_vibration_x`, `bearing_temperature_3`, `pump_outlet_pressure`): $\text{stolen}_{\text{flat}} = 0.420$, $\text{stolen}_{\text{sem-aug}} = 0.423$. ISA-95 grouping does not by itself defend against attribution-splitting, consistent with known TreeSHAP behaviour under collinearity (Lundberg et al., 2020). Two additional observations are worth noting. First, the semantic-only representation cannot be directly stress-tested in this way because it contains only component-level aggregates, not the raw sensor columns that the duplicates are injected into; the result JSON records this as a warning (`phase4_xai_metrics.json`). Second, a practical remedy that does not require changing the model is *component-level SHAP aggregation*: summing |SHAP| values within each ISA-95 component group as a post-processing step. This concentrates attribution at the physically-interpretable subsystem level and is already implicit in the sparsity analysis (Table 4), where the semantic-augmented encoding needs only 1.6% more components despite a 124% feature-space increase.

4.6. Calibration

ECE_{15}: flat 0.0016, semantic-only 0.0013, semantic-augmented 0.0013. Relative improvement -18.8%.

4.7. Cross-axis Summary

Table 5 consolidates the seven headline numbers.

4.8. Temporal Generalization (Robustness Check)

The results above use a stratified random 80/20 split. Because the dataset is a time series with strong row-to-row autocorrelation, such a split can place near-duplicate neighbours on both sides of the boundary and make *absolute* scores optimistic. We therefore re-ran the three-encoding comparison under a chronological split. A canonical last-20% holdout is degenerate here: all seven abnormal episodes fall in the first $\approx 75.5\%$ of the timeline, so the final 20% contains *zero* positives. We instead cut the timeline so that the earlier failure episodes (1–4) form the training period and the later episodes (5–7) form the test period – a genuine "predict future failures from past failures" evaluation with both classes present on each side (`poc/train_rf_temporal.py`, `results/phase2_temporal_metrics.json`).

Table 6. Temporal split (episodes 1–4 train, 5–7 test). Absolute scores collapse versus the random split (Table 1), but the *aggregate-only* encoding remains robust to cross-episode generalization.

Encoding	F1	PR-AUC	Recall	MCC
flat	0.369	0.918	0.230	0.448
semantic-only	**0.871**	**0.978**	**0.784**	**0.868**
semantic-augmented	0.381	0.936	0.242	0.449

Two effects stand out (Table 6). First, every encoding's operating-point performance drops sharply relative to the random split, confirming that the near-perfect random-split scores are optimistic for this autocorrelated series. Second, the *semantic-only* encoding is markedly more robust to cross-episode generalization (F1 0.871, recall 0.784) than either the flat (F1 0.369, recall 0.230) or the semantic-augmented (F1 0.381) encoding; the paired-bootstrap gain of semantic-only over flat is $\Delta\text{F1} = +0.50$ $[+0.49, +0.51]$, $p < 0.001$. The component aggregates (per-group mean/std/max/min/range) evidently capture a failure signature that transfers across episodes, whereas the high-cardinality raw sensors offer the greedy tree learner more episode-specific split points to overfit – which is also why *re-adding* the raw sensors (semantic-augmented) drags performance back down to the flat level. PR-AUC degrades more gracefully than F1 (0.92–0.98), indicating the ranking is partially preserved while the fixed 0.5 operating threshold is what collapses for flat and augmented. The ordering is consistent across cut fractions $0.45/0.50/0.55$ (semantic-only F1 $0.88/0.87/0.87$ vs. flat $0.35/0.37/0.34$); at a 0.60 cut all three recover to F1 ≈ 0.98 precisely because the dominant episode then leaks into training, confirming the cross-episode mechanism.

This probe should be read as exploratory, not a second headline result: the test positives are dominated by a single long recovering episode, so the estimate has high variance, and a rigorous treatment requires rolling-origin or leave-episode-out cross-validation on a benchmark with more failure events (Section 5.C). It nonetheless reframes the contribution usefully: the semantic-augmented encoding maximizes *in-distribution* explanation quality (faithfulness, stability, calibration; Tables 3–5), while the aggregate-only encoding maximizes *temporal* robustness – both vindicating the

ISA-95 semantic layer as more than an interoperability convenience.

5. DISCUSSION

5.1. Why Semantic Encoding Helps Faithfulness

TreeSHAP allocates attribution by averaging over feature orderings; under the semantic-augmented encoding many features are statistical summaries of physically-coherent groups. The model can therefore route predictive signal through aggregate features without being forced to average across near-duplicate raw sensors, sharpening the SHAP rank order and reducing the fraction of marginal attribution outside the model's actual decision logic. The deletion-AUC test exposes this: removing top-ranked features in the semantic model collapses the prediction faster.

5.2. Comparison with Prior Industrial XAI Work

Industrial-prognostics SHAP applications (Serradilla et al., 2022; Nor et al., 2021; Vollert et al., 2021) typically report SHAP plots and domain-expert review without quantifying explanation quality. Our results suggest that publishing only F1 / PR-AUC misses a $\sim 28\%$ gain in faithfulness obtainable from an information-model layer that is, on many sites, already deployed via an IIH. Conversely, IIH-focused papers (Profanter et al., 2019; Schleipen et al., 2016; Mazzetti et al., 2020; Todkar, Sarkar, & Solanki, 2025; Todkar, Sarkar, Solanki, & Tylka, 2025) stop at data integration; this work quantifies what that integration is worth at the explanation layer.

5.3. Limitations

1. PR-AUC saturation: predictive gains are bounded by the near-perfect baseline. Replication on a non-saturated dataset (PHM 2008, NASA bearing) is the obvious next step. With F1 > 0.996, the test set contains very few uncertain predictions (probability $\in [0.2, 0.8]$), precluding a difficulty-stratified XAI evaluation on this dataset; such an evaluation should be prioritised on a harder benchmark where explanation quality matters most.
2. Single architecture (RF) and single dataset.
3. One ISA-95 mapping; a sensitivity study on alternative groupings would distinguish the principle from the specific grouping.
4. Redundancy stress unaffected: TreeSHAP collinearity weakness is inherited (Section 4.D).
5. Split methodology: the headline metrics use a stratified random split, appropriate for the in-distribution comparison but optimistic for an autocorrelated series. The temporal probe of Section 4.8 is itself limited – its test positives are dominated by a single recovering episode – so the striking robustness of the aggregate-only encoding needs confirmation under rolling-origin or leave-episode-out cross-validation on a benchmark with more failure events.
6. OPC UA roundtrip fidelity: the test bench currently exhibits a residual numeric drift ($\bar{\delta} = 8.55$, $\max_c \delta_c = 79.17$; Eq. 1) rather than bit-for-bit reconstruction. We attribute this to row-alignment in the asynchronous snapshotting rather than an OPC UA protocol limitation (the wire format carries exact 64-bit floats); a deterministic-alignment fix is future work and does not affect the encoding comparison, which uses the source CSV directly.

5.4. Path to IIH Deployment

Because the asset model is the single source of truth and OPC UA companion specifications are vendor-neutral (OPC Foundation, 2022; International Electrotechnical Commission, 2020), the same `ComponentGroup` and `ISA95Path` properties populate the IIH semantic model (Siemens AG, 2024), the captured datasets, and the SHAP component-aggregator. The XAI evaluation harness is portable to a real IIH installation by changing the OPC UA endpoint URL, and to other companion specifications (PackML, Robotics) by swapping the asset-model module.

To make the benefit concrete, consider an operator receiving a fault alert. With a flat-encoded model, the SHAP report names individual sensors (*sensor_23*, *sensor_47*); the operator must manually consult a tag dictionary to identify which physical subsystem each tag belongs to before routing the work order. With the semantic-augmented model the component-level attribution (Table 4) already answers *which subsystem* – e.g., *BearingAssembly: 68% of attribution* – enabling immediate routing to the correct maintenance team without an extra lookup. The 28% faithfulness gain amplifies this advantage: features ranked as important by TreeSHAP are more likely to be genuinely decision-critical rather than incidental correlates, reducing the risk of a team investigating the wrong component.

6. CONCLUSION

This study investigated whether semantic information from OPC UA companion specifications can improve post-hoc XAI for industrial fault-prognostics models. Using an ISA-95-aware feature representation and a fixed predictor, we observed near-equivalent predictive performance ($\Delta\text{F1} \leq 0.0024$) alongside substantial improvements in explanation quality. The semantic representation reduced SHAP deletion AUC by 28%, improved Lipschitz stability by 52%, preserved component-level conciseness despite a $2.2\times$ larger feature space, and reduced calibration error by 19%.

These results demonstrate that semantic metadata already available through industrial interoperability infrastructure can produce more faithful, stable, and reliable explanations with-

out additional manual asset modeling. Although evaluated using ISA-95 and Siemens IIH, the proposed methodology is applicable to other OPC UA companion specifications and production endpoints. Code, data, and results are released to support reproducibility.

ACKNOWLEDGMENT

The authors thank the Siemens Foundational Technology team and the SIMATIC IIH product group for technical discussions on the IIH semantic model and OPC UA Aggregation server.

NOMENCLATURE

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in neural information processing systems (neurips)* (Vol. 31).

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. In *Proceedings of the 29th international joint conference on artificial intelligence (ijcai)*.

Elgendy, S. (2018). *Pump sensor data — time-series analysis.* Kaggle dataset and analysis notebook. Retrieved from `https://www.kaggle.com/code/shawkyelgendy/pump-sensor-data-timeseriesanalysis` (Accessed 2026-06-13)

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning (icml)* (pp. 1321–1330).

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., ... Höhne, M. M.-C. (2023). Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, *24*, 1–11.

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In *Advances in neural information processing systems (neurips)* (Vol. 32).

International Electrotechnical Commission. (2013). *IEC 62264: Enterprise-Control System Integration.*

International Electrotechnical Commission. (2020). *IEC 62541: OPC Unified Architecture.*

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems (neurips)* (Vol. 30, pp. 4765–4774).

Mazzetti, M., et al. (2020). OPC UA and Industry 4.0: Enabling interoperability for IIoT data integration. *IEEE Industrial Electronics Magazine*.

Nor, A. K. M., Pedapati, S. R., Muhammad, M., & Leiva, V. (2021). Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses. *Sensors*, *21*(23), 8020.

OPC Foundation. (2022). *OPC 10030: ISA-95 Common Object Model Companion Specification, v4.2.* Retrieved from `https://reference.opcfoundation.org/specs/OPC-10030/4.2`

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the british machine vision conference (bmvc)*.

Profanter, S., Tekat, A., Dorofeev, K., Rickert, M., & Knoll, A. (2019). OPC UA versus ROS, DDS, and MQTT: Performance evaluation of Industry 4.0 protocols. In *Proceedings of the ieee international conference on industrial technology (icit)* (pp. 955–962).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the aaai conference on artificial intelligence* (pp. 1527–1535).

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673.

Schleipen, M., Gilani, S.-S., Bischoff, T., & Pfrommer, J. (2016). OPC UA & Industrie 4.0: Enabling technology with high diversity and variability. *Procedia CIRP*, *57*, 315–320.

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. *Applied Intelligence*, *52*, 10934–10964.

Siemens AG. (2024). *IIH Semantics: Semantic data modelling on the industrial edge.* Product documentation.

Siemens AG. (2025). *SIMATIC Industrial Information*

Hub (IIH). SIMATIC Industrial Edge product page. Retrieved from https://www.siemens.com/en-us/products/simatic-apps/industrial-information-hub/

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods. In *Proceedings of the aaai/acm conference on ai, ethics, and society (aies)* (pp. 180–186).

Todkar, A., Sarkar, M., & Solanki, J. (2025, October). Semantic framework for IT-OT integration in industrial environments. In *Annual conference of the phm society* (Vol. 17). doi: 10.36001/phmconf.2025.v17i1.4551

Todkar, A., Sarkar, M., Solanki, J., & Tylka, J. (2025). Approaches to automatic discovery and modeling of industrial assets for IT/OT integration. In *Proceedings of the 2025 ieee 21st international conference on automation science and engineering (case)* (pp. 2341–2346). doi: 10.1109/CASE58245.2025.11164146

Vollert, S., Atzmueller, M., & Theissler, A. (2021). Interpretable machine learning: A brief survey from the predictive maintenance perspective. In *Proceedings of the ieee international conference on emerging technologies and factory automation (etfa).*

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. (2019). On the (in)fidelity and sensitivity of explanations. In *Advances in neural information processing systems (neurips)* (Vol. 32).

Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, *13*(3), 2213–2227.