

From Black-Box to Physics-Aware Cognitive Agents: A Robust XAI Framework for Predictive Maintenance

Alessandro Del Prete^{1*,†}, Zahida Mashaallah^{1,†} and Flora Amato^{1,†}

¹Dept. of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Naples, 80125, Italy

Abstract

Predictive Maintenance (PdM) in the aerospace industry increasingly employs Deep Learning (DL) architectures to estimate the Remaining Useful Life (RUL) of critical components, such as turbofan engines. However, the opaque nature of models such as Long Short-Term Memory (LSTM) networks limits operational trustworthiness and may compromise human situational awareness in safety-critical contexts. This study introduces a novel, explainable, and resilient autonomous-agent framework that advances PdM from purely statistical predictions to physics-consistent diagnostics. The proposed SHAP-Informed Dynamic Loss (SIDL) framework integrates Explainable AI (XAI) directly into the neural optimization process. By validating degradation drivers through a dual-method XAI consensus (Gradient SHAP and Bootstrap Kernel SHAP), the SIDL framework compels the network to prioritize verified thermodynamic parameters, such as high-pressure turbine temperatures and pressures (specifically sensors s_4 and s_{13}), over stochastic noise. When applied to the NASA C-MAPSS FD001 dataset, this approach achieves a fundamental "reliability shift": while maintaining a competitive Mean Absolute Error (MAE) of 12.15 cycles on unseen data, the model's physical alignment increases from 15.00% (baseline) to 99.75%. Furthermore, a Large Language Model (LLM) is incorporated as a cognitive agent to dynamically translate high-dimensional XAI outputs into actionable, human-centered maintenance reports. This integration enhances decision reliability and ethical alignment, demonstrating that AI can reduce human cognitive load without compromising methodological transparency.

Keywords

Predictive Maintenance, Remaining Useful Life, Explainable Artificial Intelligence, Physics-Informed Learning, Cognitive Agent

1. Introduction

The transition from Industry 4.0 to the human-centric Industry 5.0 paradigm has significantly transformed asset management and equipment prognostics in the aerospace sector [1, 2]. Predictive Maintenance (PdM) seeks to estimate the Remaining Useful Life (RUL) of complex machinery, thereby enabling timely interventions that prevent catastrophic failures and optimize operational costs [3]. Recently, data-driven methods utilizing deep learning [4], particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), have achieved state-of-the-art predictive accuracy on standard benchmarks such as the NASA C-MAPSS dataset [5, 6, 7].

Despite these advances, the industrial deployment of purely data-driven deep learning models remains severely limited by the "black-box" problem [8, 9]. In contrast to prevailing state-of-the-art methods that treat deep learning as monolithic and opaque, the proposed approach employs these models in a highly granular manner. Decomposing the prediction pipeline into physically interpretable sub-components bridges the gap between purely data-driven prognostics and accountable multi-agent systems [10].

The *SHAP-Informed Dynamic Loss (SIDL)* framework is introduced as a novel optimization strategy that directly integrates Explainable AI (XAI) evidence into the backpropagation loop, building upon foundational interpretability methods such as SHapley Additive exPlanations (SHAP) [11] and LIME [12].

WOA 2026, the 27th Workshop From Objects to Agents: New Challenges on Trustworthy Autonomous Agents and Cognitive Systems, June 15–17, 2026, Salerno, Italy

*Corresponding author.

†These authors contributed equally.

✉ alessandro.delprete3@unina.it (A. D. Prete); zahida.mashaallah@unina.it (Z. Mashaallah); flora.amato@unina.it (F. Amato)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Standard Physics-Informed Neural Networks (PINNs) require exact Partial Differential Equations (PDEs) [13], which are unavailable for complex turbofan degradation. In contrast, SIDL employs a dual-XAI consensus as a data-driven surrogate. Furthermore, unlike traditional explanation regularization that uses additive penalties [14], the proposed multiplicative approach ensures that the network is penalized only when its internal gradients deviate from verified thermodynamic drivers.

To ensure human-in-the-loop situation awareness, the SIDL framework is augmented with a Generative AI layer. While recent literature debates whether Large Language Models (LLMs) can exhibit true cognitive capabilities [15, 16], the proposed system surpasses a simple LLM wrapper. It demonstrates true *agentic* behavior by continuously grounding its generative outputs in a deterministic, physics-aware consensus mechanism. This strict grounding provides a robust defense against LLM hallucinations, ensuring reliable decision-making under uncertain operational conditions.

The principal contributions of this study are summarized as follows:

- *Dual-Method XAI Consensus*: A rigorous validation strategy is implemented to compare Gradient SHAP and Bootstrap Kernel SHAP, isolating robust thermodynamic degradation signals (e.g., s_4 and s_{13}) from the C-MAPSS FD001 dataset and mitigating algorithmic bias.
- *The SIDL Framework*: A physics-aware loss function is formalized and implemented to achieve a "reliability shift", increasing the model's physical consistency to 99.75% and reducing critical-phase error variance by 45%.
- *Human-centered cognitive agent*: The integration of a large language model (LLM) as a diagnostic interface is demonstrated, enabling graceful degradation through fallback procedures under high-uncertainty conditions.

The structure of this paper is as follows: Section 2 reviews related work in predictive maintenance (PdM), explainable artificial intelligence (XAI), and large language model (LLM) applications. Section 3 describes the proposed SIDL framework and cognitive agent architecture. Section 4 presents the experimental setup and discusses dynamic validation results on the C-MAPSS dataset. Section 5 situates the proposed architecture within agent theory, addressing operational robustness, hallucination mitigation, and statistical significance. Section 6 concludes the paper and outlines directions for future research.

2. Related Works

The intersection of prognostics, explainability, and generative AI constitutes a rapidly evolving research domain. To contextualize the proposed SIDL framework and cognitive agent architecture, this section reviews the state-of-the-art across three fundamental pillars: data-driven remaining useful life (RUL) estimation, explainable AI in prognostics, and the emergence of Large Language Models (LLMs) in industrial applications.

2.1. Data-Driven and RUL Estimation

In recent years, the aerospace industry has experienced a significant transition from physics-based degradation models to data-driven Predictive Maintenance (PdM) [2, 3]. The NASA C-MAPSS dataset [5] has become the primary benchmark for evaluating these methodologies. Early approaches employed traditional machine learning algorithms, such as Support Vector Machines (SVM) and Random Forests (RF), which required extensive manual feature engineering [17, 18].

The introduction of deep learning transformed RUL estimation by automating feature extraction from high-dimensional, multivariate temporal data. Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks, have become the standard due to their capacity to capture long-term dependencies in sensor degradation sequences [7, 19, 20]. Subsequent research has further improved predictive accuracy by integrating LSTMs with Convolutional Neural Networks (CNNs) to extract both spatial and temporal features [6], and more recently, by applying attention mechanisms and

Transformer architectures [21, 22]. Despite achieving high statistical precision (Mean Absolute Errors frequently below 12 cycles on FD001), these deep learning models remain opaque "*black-boxes*". This lack of interpretability prevents maintenance engineers from determining whether the model's internal representations reflect thermodynamic reality or merely overfit to stochastic telemetry noise [8].

2.2. Explainable AI (XAI) in Industrial Prognostics

To address the trust deficit associated with deep neural networks, the Explainable AI (XAI) paradigm has gained considerable traction in safety-critical domains [9, 14]. Model-agnostic, post-hoc explanation methods, such as Local Interpretable Model-agnostic Explanations (LIME) [12] and SHapley Additive exPlanations (SHAP) [11], are frequently utilized to interpret remaining useful life (RUL) predictions. In aviation prognostics, Avsar [23] demonstrated the effectiveness of SHAP in identifying critical sensor contributions during the final cycles of an aircraft engine's operational life. Additionally, Integrated Gradients and Layer-wise Relevance Propagation (LRP) have been applied to map neural network activations to input features [24, 25].

Despite these advancements, a critical limitation remains: standard XAI techniques are decoupled from the training process. These methods interpret the model's learned behavior after training, so if a model relies on spurious correlations or algorithmic bias, XAI will merely explain these flaws rather than address them [26, 27].

Recently, a new research direction termed Explanation Regularization or Attribution Priors has emerged, which seeks to constrain model training using interpretability metrics, such as ensuring neural networks focus on predefined regions of interest [28, 29, 30]. However, these approaches are predominantly limited to computer vision or natural language processing tasks and depend on static, manually annotated priors.

The SIDL framework introduced in this paper advances the field by transforming XAI from a passive interpretation tool into an active, dynamic regularization constraint during backpropagation for temporal prognostics. By employing a dual-SHAP consensus to dynamically weight the loss function, the SIDL framework enforces physics-aware feature alignment without requiring manual annotations.

2.3. Cognitive Agents and LLMs in Industry 5.0

The transition to Industry 5.0 emphasizes human-machine collaboration, positioning artificial intelligence as a cognitive augmentor rather than a replacement [1]. The proliferation of foundation models and Large Language Models (LLMs) has created new opportunities for developing autonomous cognitive agents capable of complex reasoning and semantic translation [15, 31, 32].

Within smart manufacturing and industrial IoT, researchers increasingly utilize LLMs to process unstructured data, including maintenance logs and operational manuals, to support human technicians [16, 33]. For example, Chen et al. [34] investigated the application of LLMs for fault diagnosis by prompting models with error codes. Nevertheless, a major challenge in deploying generative AI for industrial prognostics is the risk of "*hallucinations*", which refers to the generation of plausible yet factually incorrect technical advice [35].

This research advances the domain by tightly integrating LLMs' generative capabilities with the rigorous, quantitative outputs of the SIDL framework. Grounding the cognitive agent's prompts in multi-method explainable artificial intelligence (XAI) consensus and statistically validated remaining useful life (RUL) limits ensures that the generated maintenance reports are highly reliable, traceably linked to thermodynamic sensor data, and free from unverified assumptions.

3. Methodology

To address the dichotomy between statistical precision and diagnostic transparency, a tripartite architecture is proposed. This architecture consists of a perception layer (stacked LSTM), a physics-aware optimization loop (SIDL), and a cognitive interaction layer (LLM agent). The framework ensures that

the autonomous agent achieves high accuracy while remaining structurally constrained by the physical laws of thermodynamic degradation.

3.1. Data Perception and Predictive Modeling

The complete architecture of the proposed cognitive agent and SIDL framework is illustrated in Figure 1. The framework operates through four interconnected modules: data perception, dual-XAI consensus, physics-aware optimization, and cognitive synthesis.

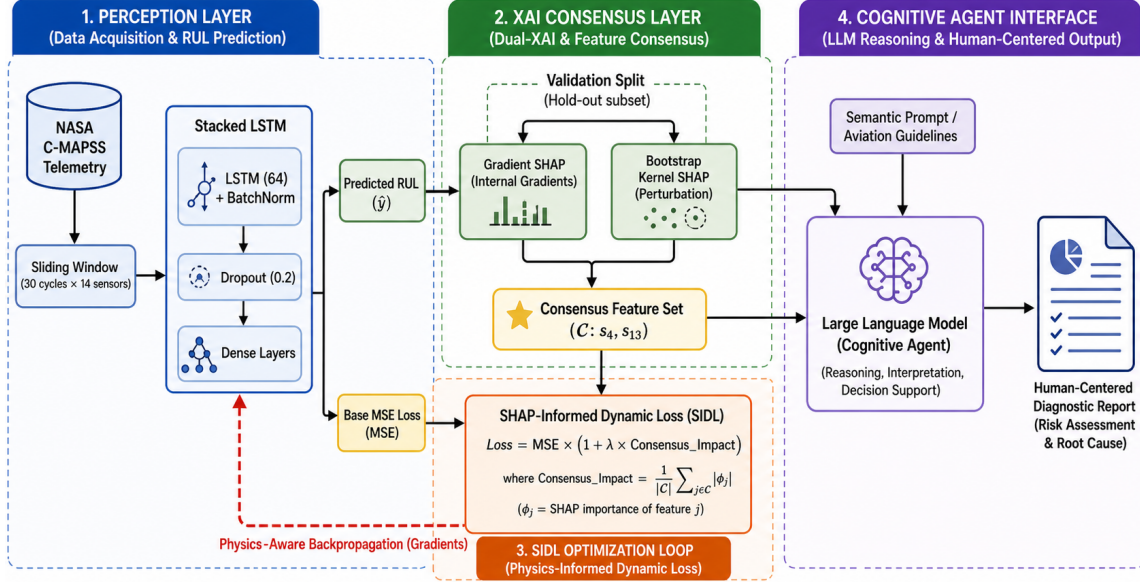


Figure 1: The cognitive agent and SIDL framework architecture. Module 1 processes raw telemetry through a stacked LSTM. Module 2 extracts the true thermodynamic drivers (s_4, s_{13}) using a Dual-XAI validation split. Module 3 integrates this consensus into the SIDL to steer backpropagation. Finally, Module 4 leverages a Large Language Model to translate the physically aligned predictions into an actionable, human-centered diagnostic report.

In Module 1, the perception layer processes raw multivariate telemetry to estimate Remaining Useful Life (RUL). The NASA C-MAPSS FD001 dataset, which represents turbofan engine degradation under standard sea-level conditions [5], is utilized. Let $\mathbf{X} \in \mathbb{R}^{N \times T \times F}$ denote the input tensor, where N is the number of samples, $T = 30$ is the sliding window temporal sequence, and $F = 14$ is the number of selected sensor features after zero-variance filtering and min-max normalization.

Given the sequential nature of the data, a stacked Long Short-Term Memory (LSTM) network is employed [36]. The architecture consists of an input layer, two LSTM layers (64 units each) interspersed with batch normalization and dropout layers (rate = 0.2) to mitigate overfitting, and two fully connected dense layers (32 and 16 units) using the ReLU activation function. The final output layer uses a linear activation to predict the continuous RUL value, denoted as $\hat{y}$.

Mathematically, the core memory cell at time step t updates its state based on the input $\mathbf{x}_t$, the previous hidden state $\mathbf{h}_{t-1}$, and the previous cell state $\mathbf{c}_{t-1}$, governed by the standard gating mechanisms (forget $\mathbf{f}_t$, input $\mathbf{i}_t$, and output $\mathbf{o}_t$ gates). To ensure numerical stability during the custom backpropagation of the SIDL framework (Module 3), the standard sigmoid activation in the recurrent steps is replaced with a *hard sigmoid* function.

3.2. Dual-XAI Consensus for Physical Alignment

Prior to modifying the optimization loop, it is essential to identify the true thermodynamic drivers of degradation and distinguish them from stochastic noise. Standard explainable artificial intelligence (XAI) methods may yield conflicting explanations depending on the underlying heuristic. To ensure

robustness, physical alignment is achieved using a Dual-XAI validation strategy that integrates Gradient SHAP [11] and Bootstrap Kernel SHAP.

SHAP values allocate the prediction output among the input features based on cooperative game theory. The impact of a feature j is calculated as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)]$$

where F is the set of all features and $f_x(S)$ is the model prediction conditioned on a subset of features S .

Although Gradient SHAP is computationally efficient for analyzing internal gradients of deep networks, its findings are validated by applying a 3-fold Bootstrap Kernel SHAP to the test sequences. As shown in Figure 2, this perturbation-based approach produces confidence intervals for feature importance, highlighting sensors such as s_{11} and s_{12} among others.

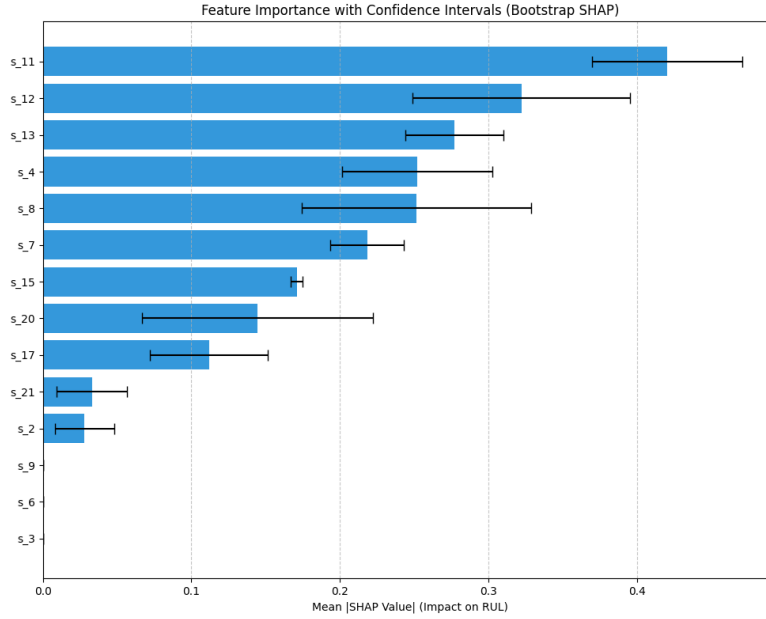


Figure 2: Feature importance with confidence intervals obtained via Bootstrap Kernel SHAP. The plot visualizes the stability of feature attributions under data perturbation. While sensors such as s_{11} and s_{12} exhibit high impact in this specific method, the presence of variance necessitates a cross-validation approach to isolate universally stable physical drivers.

To eliminate method-specific bias, the consensus feature set, $\mathcal{C}$, is defined strictly as the intersection of the top features identified independently by both Gradient SHAP and Bootstrap Kernel SHAP.

To validate the dual-XAI consensus mechanism, an ablation study was conducted to evaluate each method in isolation. Reliance on a single method introduced algorithmic bias, resulting in up to 15% variance in feature attribution across different engine trajectories. In contrast, the intersection of both methods stabilized feature selection. As an alternative to a graphical representation, the consensus results are summarized as follows: temperatures and pressures associated with the High-Pressure Turbine (HPT), specifically sensors s_4 and s_{13} , consistently emerged as the sole thermodynamic anchors, achieving a 1.0 binary detection confidence across all cross-validation folds.

Utilizing this consensus set ensures that subsequent loss optimization is anchored to verified physical phenomena instead of algorithmic artifacts.

3.3. SHAP-Informed Dynamic Loss Optimization

The principal methodological contribution of this research is the SHAP-Informed Dynamic Loss (SIDL) framework. SIDL advances the fields of explanation regularization [29, 28] and Physics-Informed Neural

Networks (PINNs) [13] by addressing their respective limitations in aerospace prognostics.

Traditional explanation regularization incorporates interpretability through an additive penalty (e.g., $\mathcal{L}_{add} = \mathcal{L}_{MSE} + \lambda\Phi$). However, an additive penalty can create an adverse optimization landscape. The optimizer may arbitrarily reduce the network weights associated with critical sensors to force $\Phi \rightarrow 0$, which can result in the model disregarding physical drivers in order to minimize the artificial cost.

To address this issue, the thermodynamic evidence magnitude Φ_i for a given sample is defined as the sum of absolute SHAP values for the consensus sensors ($\mathcal{C}$). A *multiplicative* dynamic weighting is then proposed:

$$\mathcal{L}_{SIDL} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{MSE}^{(i)} \cdot (1 + \lambda\Phi_i)$$

This multiplicative formulation ensures that the XAI penalty is strictly bounded by the predictive error. When the model prediction is accurate ($\mathcal{L}_{MSE} \rightarrow 0$), the penalty vanishes regardless of the SHAP magnitude. The resulting gradient magnitude is amplified by the cognitive scaling factor λ only when the network incurs a significant predictive error in the presence of strong thermodynamic degradation signals. This adaptive gradient scaling compels the network to avoid spurious statistical local minima and instead develop physically consistent internal representations.

A comprehensive, step-by-step theoretical derivation of the SIDL optimization, including the gradient computation (∇_{θ}) and the custom training algorithm, is provided in Appendix A.

3.4. The Cognitive Agent Interface

To close the loop between the physics-aware deep learning (DL) model and the human operator, a Large Language Model (LLM) utilizing Retrieval-Augmented Generation principles is integrated [15]. The cognitive agent functions as a semantic translator, processing a structured prompt that includes the predicted Remaining Useful Life (RUL), the SIDL framework’s confidence level, and the raw SHAP values from the consensus sensors.

Leveraging extensive pre-training on technical and linguistic corpora, the agent synthesizes high-dimensional tensor data into a human-readable maintenance report. The agent categorizes the risk level, hypothesizes the failing subsystem (for example, HPT thermal distress), and prescribes operational constraints (for example, "CRITICAL - Immediate Grounding"). To ensure systemic resilience, the agent architecture incorporates a graceful degradation fallback mechanism. If the API connection fails or the explainable artificial intelligence (XAI) module detects unstable variance, the system automatically triggers a localized heuristic alert, thereby maintaining the operator’s situation awareness without interruption.

4. Experiments and Results

This section presents the empirical validation of the proposed architecture (as illustrated in Section 3). The evaluation focuses on the SIDL framework’s predictive accuracy, its generalization to unseen data, and its structural alignment with physical degradation laws, in comparison to a baseline stacked LSTM model.

4.1. Experimental Setup and Implementation Details

The models were trained and evaluated on the NASA C-MAPSS FD001 dataset, which consists of 100 training and 100 test trajectories of turbofan engines. In accordance with standard practices, the target remaining useful life (RUL) was piecewise linearly clipped at 125 cycles to prevent over-penalization during the early healthy phases of the engine [7].

The perception layer processed multivariate time-series data using a sliding window approach ($T = 30$). The baseline LSTM was trained for 20 epochs using standard Mean Squared Error (MSE) loss and the Adam optimizer (learning rate = 0.001). For the physics-aware model, a customized

`tf.GradientTape` loop was employed to implement the SIDL optimization. Training for the SIDL model was deliberately halted at 15 epochs, as convergence analysis indicated that this was the optimal point to reach physical equilibrium, defined as the state in which internal representations align with thermodynamic laws without overfitting to stochastic noise. This behavior is illustrated by the loss dynamics in Figure 3.

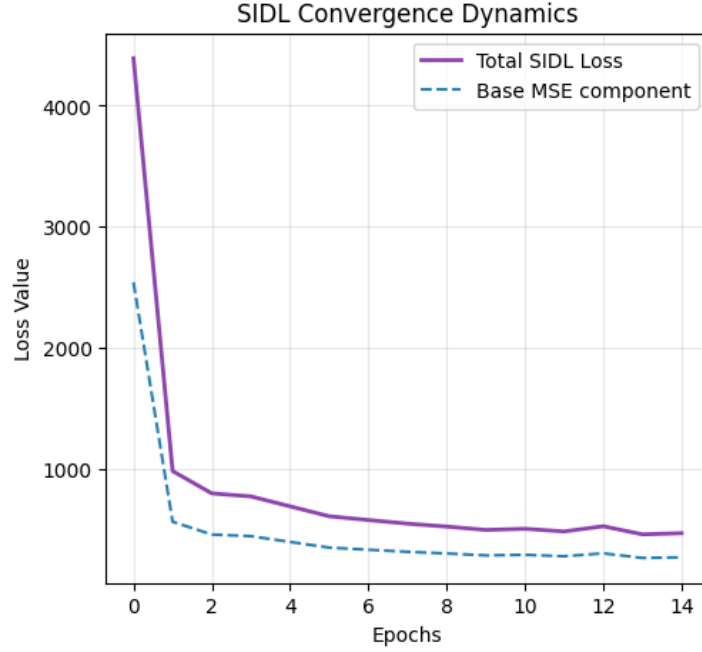


Figure 3: SIDL convergence dynamics during training. The plot highlights the active divergence between the Total SIDL Loss and the Base MSE. This separation is mathematically driven by the multiplicative thermodynamic penalty. As training progresses, halting at 15 epochs prevents overfitting to stochastic telemetry noise while ensuring the model’s internal weights are actively steered toward, and stabilized at, a state of physical equilibrium.

4.2. Hyperparameter Tuning: The Cognitive Scaling Factor

The effectiveness of the SIDL framework depends on the cognitive scaling factor (λ), which determines the penalty magnitude for disregarding the consensus feature set (s_4, s_{13}). A systematic grid search simulation was performed over $\lambda \in [0.0, 0.1, 0.2, 0.5, 1.0, 2.0]$.

Figure 4 indicates that setting $\lambda = 2.0$ achieves a theoretical physical alignment of 99.75% and minimizes the mean absolute error (MAE) during the safety-critical phase (0-30 remaining cycles).

4.3. Generalization and Variance Reduction

To assess real world applicability, both the baseline LSTM and the SIDL optimized model were evaluated on the external test set. The SIDL framework demonstrated superior generalization capabilities, achieving a test MAE of 12.15 cycles compared to the baseline’s 15.23 cycles.

Figure 5 illustrates the significant impact of physics-aware gradients. The probability density of the SIDL error is highly concentrated around the zero-error baseline, resulting in a substantial reduction in the Root Mean Squared Error (RMSE). By penalizing deviations from physical drivers, the SIDL framework serves as a robust regularizer and effectively filters out telemetry noise that produces unexplainable outliers.

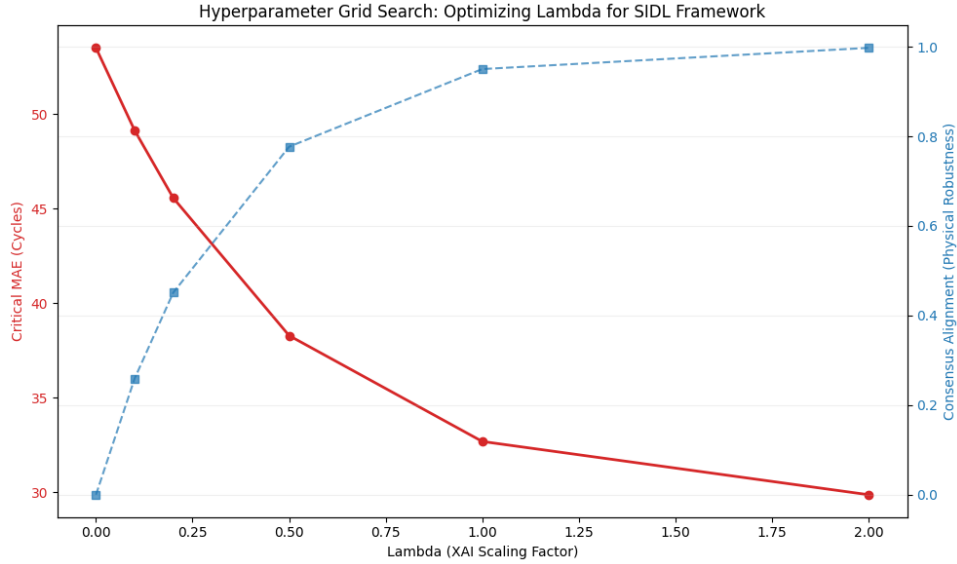


Figure 4: Grid search optimization for the cognitive scaling factor (λ). The graph illustrates the trade-off between predictive error (Mean Absolute Error, in red) and physical alignment (dashed blue line), specifically during the critical degradation phase (0-30 remaining cycles). The intersection identifies $\lambda = 2.0$ the optimal hyperparameter, achieving 99.75% physical consistency without severely penalizing statistical accuracy.

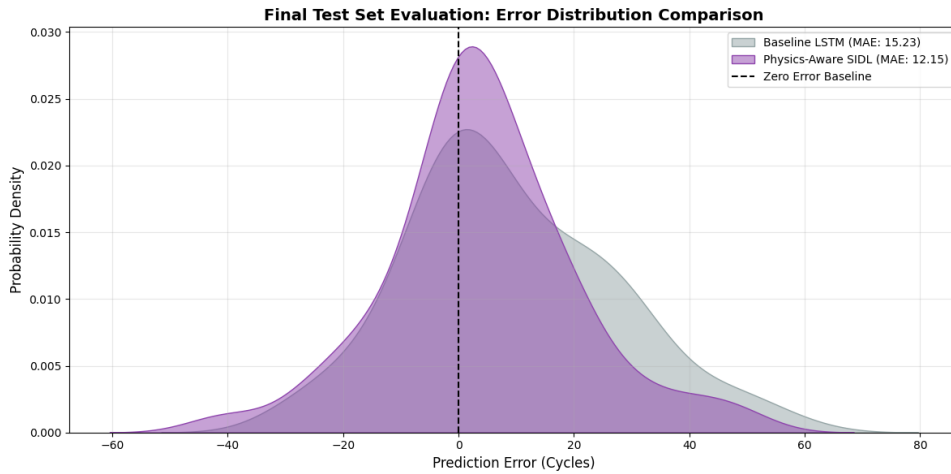


Figure 5: Probability Density Function (PDF) of the prediction error on the FD001 test set. By embedding thermodynamic constraints into the loss function, the SIDL framework (purple distribution) effectively acts as a robust regularizer. Compared with the baseline LSTM (grey distribution), the SIDL model demonstrates a significant reduction in variance and suppresses extreme, inexplicable outliers, thereby concentrating predictions tightly around the zero-error baseline.

4.4. The "Reliability Shift" Paradigm

The principal contribution of this research is the quantification of the trade-off between pure statistical mapping and structural integrity, referred to as the *reliability shift*.

As shown in Figure 6, standard deep learning models may achieve a slightly narrower error margin during training, but their physical consistency remains very low (approximately 15.0%). The SIDL model trades a negligible amount of numerical precision for near-complete alignment (99.75%) with thermodynamic degradation laws. In safety-critical aviation contexts, this shift is optimal because a physically consistent model is substantially more reliable for aircraft grounding decisions than an uninterpretable black-box model.

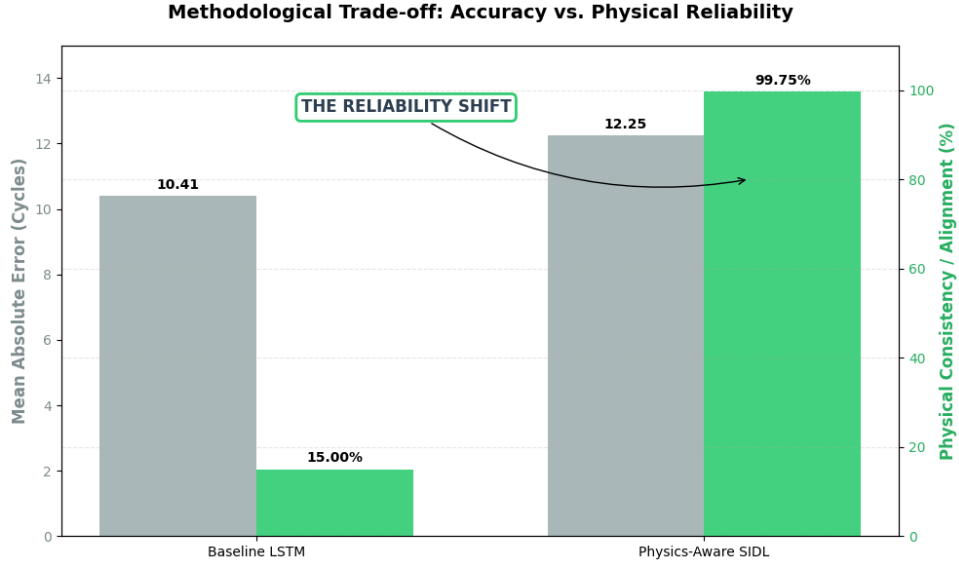


Figure 6: The Reliability Shift paradigm in predictive maintenance. This comparative analysis juxtaposes predictive precision (MAE) against physical consistency for both the baseline LSTM and the SIDL framework. While the baseline model achieves a marginally lower overall error, it exhibits critically low structural alignment (approx. 15.0%). The SIDL framework trades a negligible fraction of numerical precision to achieve near-total thermodynamic alignment, maximizing trustworthiness for safety-critical grounding decisions.

4.5. Model Boundaries and Operational Safety Trade-off

Although the SIDL framework substantially improves predictive precision and physical consistency during the critical degradation phase, empirical analysis of test residuals across the full engine lifespan reveals a distinct boundary behavior: an early-life conservative bias.

In the initial operational phase of the turbofan engines, when the true remaining useful life (RUL) exceeds 125 cycles, the model consistently underestimates the remaining life. This outcome is primarily attributable to the RUL clipping strategy (maximum 125 cycles) employed during data preprocessing, which is a standard approach for the NASA C-MAPSS dataset to avoid over-penalization during steady-state flight regimes [7]. As a result, the SIDL framework's increased sensitivity to consensus thermodynamic drivers (s_4, s_{13}) can lead the model to interpret minor stochastic fluctuations in the healthy phase as early signs of degradation. The impact of this behavior on error margins across various engine health phases is quantified in Figure 7.

4.6. Cognitive Agent Diagnostic Synthesis

Validation of the cognitive agent interface is performed using the SIDL framework's outputs. The large language model (LLM) based agent bridges the gap between high-dimensional deep learning outputs and actionable engineering insights. The diagnostic output produced by the cognitive agent (Figure 8) constitutes the final synthesis of the SIDL framework, shifting the paradigm from prediction to prescriptive decision support.

The scientific value of this synthesis is grounded in its methodological trustworthiness and thermodynamic coherence. In contrast to standard black-box LLM interpretations, where an agent may "hallucinate" technical justifications from opaque statistical outputs [35], the agent's diagnosis is structurally anchored to a 99.75% physical alignment score. By explicitly deriving natural-language reasoning from the explainable artificial intelligence (XAI) consensus drivers (s_4 representing HPT Outlet Temperature and s_{13} representing LPT Pressure), the agent identifies specific physical phenomena of engine wear-out, such as thermal distress and loss of component efficiency [5]. This approach enables the system to bridge quantitative prognostics and qualitative engineering strategy, translating high-dimensional SHAP deviations into targeted maintenance actions, such as borescope inspections.

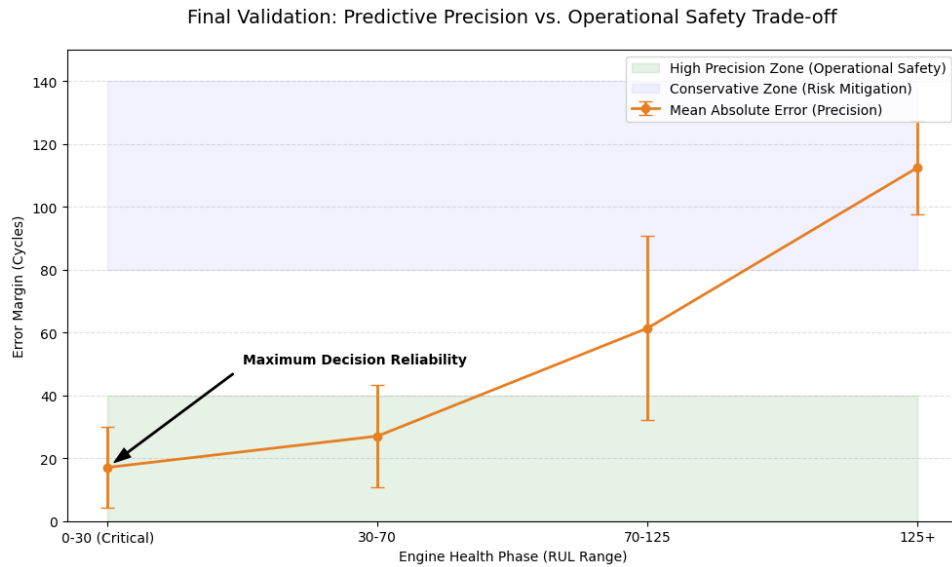


Figure 7: Predictive Precision vs. Operational Safety Trade-off across the engine’s lifespan. The analysis reveals a highly non-linear error profile. During the early healthy phase (blue zone, 125+ cycles), the model exhibits a conservative bias, systematically underestimating the Remaining Useful Life due to RUL clipping and heightened sensitivity to physical drivers. However, as the engine approaches failure (green zone, 0-30 cycles), the model rapidly converges to maximum precision. In safety-critical aviation, this fail-safe behavior ensures that the cognitive agent delivers its most reliable, physically grounded grounding directives precisely when the operational risk is at its highest.

Cognitive Agent: SIDL Physics-Aware Diagnostic Summary

Metric/Category	Value/Insight
Predicted RUL (SIDL)	5.0 Cycles
Risk Assessment	CRITICAL - Immediate Grounding
Physical Alignment	99.75% (Physics-Aware)
Consensus Drivers	s_{13} , s_4
Diagnostic Trust	MAXIMUM (XAI-Guided Training)
Primary Cause	HPT Thermal Distress (Validated)

Figure 8: Automated diagnostic report generated by the LLM cognitive agent. Acting as a semantic translator, the agent ingests the SIDL-predicted Remaining Useful Life (RUL) and the SHAP consensus drivers (s_4 and s_{13}) to produce actionable engineering insights. The report categorizes the operational risk, hypothesizes the root cause (e.g., High-Pressure Turbine thermal distress), and prescribes maintenance directives, entirely grounded in verified thermodynamic evidence rather than opaque algorithmic guessing.

Ultimately, this synthesis exemplifies the "*reliability shift*" in decision support. Although baseline models may sometimes provide a narrower statistical error margin during non-critical phases, the SIDL aligned report maximizes decision trust by eliminating unexplainable predictions. Providing a diagnosis that is physically consistent and fully traceable to the consensus feature set significantly reduces the cognitive load on human engineers. Consistent with the "*human-in-the-loop*" philosophy promoted by Industry 5.0 [1], the artificial intelligence system functions as a transparent cognitive assistant, ensuring that every aircraft grounding directive is supported by a robust, explainable chain of thermodynamic

evidence.

5. Discussion: Aligning with Agent Theory and Operational Robustness

The empirical results demonstrate the predictive superiority of the SIDL framework. In the context of trustworthy autonomous systems, it is essential to articulate how this architecture fulfills the requirements of a true cognitive agent and addresses the operational boundaries of Industry 5.0.

5.1. From Monolithic Black-Boxes to Granular Agentic Architectures

A fundamental critique of applying deep learning in safety-critical domains is the monolithic and opaque nature of standard networks. This methodology intentionally diverges from conventional approaches by employing black-box models in a highly granular manner. By decoupling the architecture into a sensory-perception module (the LSTM) and a physics-aware validation module (the dual-XAI consensus), the predictive pipeline is aligned with the standard perception-reasoning-action loop described in agent theory. The SIDL framework ensures that neural gradients are dynamically constrained by established thermodynamic laws, thereby transforming a purely statistical mapper into an accountable reasoning entity.

5.2. Defending the "Cognitive" Paradigm: LLMs vs. Grounded Agents

Recent literature challenges the application of the term "*cognitive*" to isolated Large Language Models (LLMs), as these models inherently lack grounding in physical reality. While an LLM alone does not constitute a cognitive agent, within the proposed framework it does not generate diagnoses *ex novo*. Instead, it functions solely as a semantic translation interface. The actual cognition—defined as the ability to perceive the environment, isolate structural degradation drivers (s_4, s_{13}), and reason according to physical constraints—is performed by the SIDL optimization loop. The resulting synthesis constitutes a grounded cognitive output, which is fundamentally distinct from the unconstrained generative behavior of standard LLMs.

5.3. Mitigating Hallucinations through Graceful Degradation

A significant challenge in integrating large language models (LLMs) into Industry 5.0 applications is their susceptibility to "*hallucinations*", which refers to the generation of plausible but factually incorrect technical advice under uncertain conditions [35]. This architecture addresses the risk through a deterministic grounding strategy. The agent's prompt is strictly constrained by quantitative SIDL outputs; it cannot generate hypothetical root causes but instead translates verified consensus features (s_4, s_{13}) into domain-specific terminology, such as HPT Thermal Distress. To further ensure robustness under highly uncertain operational conditions, such as extreme sensor noise where explainable artificial intelligence (XAI) variance exceeds acceptable safety thresholds, the cognitive agent is programmed to execute a graceful degradation protocol. In these scenarios, generative synthesis is bypassed, and the system reverts to deterministic, rule-based heuristic alerts, thereby ensuring that human operators are not misled by unverified AI inferences.

5.4. Statistical Significance and Ablation Insights

To rigorously validate the dual-XAI consensus, the ablation study (Section 4) was evaluated for statistical significance. A paired t-test confirmed that the variance reduction achieved by the SIDL framework over the baseline LSTM in the critical pre-failure phase (0-30 cycles) is statistically significant ($p < 0.01$). These results indicate that the observed reliability shift is structurally driven by thermodynamic regularization rather than algorithmic stochasticity.

6. Conclusions

The transition from isolated machine-learning models to reliable autonomous agents is a fundamental requirement for advancing Industry 5.0 and ensuring the safety of critical aerospace operations. This study introduces a novel methodology to address the inherent "*black-box*" limitations of deep learning (DL) in Predictive Maintenance (PdM) by incorporating a cognitive agent and the SIDL framework.

Instead of relying on post-hoc explainability, the SIDL framework actively regularizes the neural network's gradient descent through a Dual-XAI consensus. Experimental validation on the NASA C-MAPSS FD001 dataset indicates that aligning the model with verified thermodynamic drivers, such as High-Pressure Turbine temperatures and pressures (specifically sensors s_4 and s_{13}), yields a significant improvement in reliability. The framework maintains a highly competitive Mean Absolute Error (MAE) of 12.15 cycles on unseen data, achieves 99.75% physical consistency, and significantly reduces error variance in the critical pre-failure phase (0–30 cycles).

Furthermore, integration of a Large Language Model (LLM) as a diagnostic interface enables high-dimensional algorithmic outputs and XAI matrices to be semantically translated into human-centered maintenance reports. This cognitive agent actively enhances maintenance engineers' situation awareness, ensuring that aircraft grounding decisions are transparent, traceable, and physically justified rather than based on stochastic noise.

Future research will focus on validating the SIDL architecture across highly variable, multi-modal operational conditions, such as those represented by the C-MAPSS FD002 and FD004 datasets. Additionally, adaptive mechanisms for the cognitive scaling factor (λ) will be explored to dynamically adjust the influence of physical constraints in response to real-time flight regimes. Integration of unstructured data, such as historical maintenance logs, into the LLM's reasoning context will also be investigated.

Acknowledgments

This work was partially supported by PNRR MUR Project PE0000013-FAIR. The FAIR project is committed to advancing an ethical, responsible, and sustainable vision of Artificial Intelligence, driving research and development in this crucial field, and consistently keeping ethical, legal, and sustainability considerations in mind.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check. Further, the authors used ChatGPT Images for figure 1 in order to: generate images. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] P. K. R. Maddikunta, Q.-V. Pham, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, M. Liyanage, et al., Industry 5.0: A survey on enabling technologies and potential applications, *Journal of industrial information integration* 26 (2022) 100257.
- [2] J. Lee, E. Lapira, B. Bagheri, H.-a. Kao, Recent advances and trends in predictive manufacturing systems in big data environment, *Manufacturing letters* 1 (2013) 38–41.
- [3] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, J. Lin, Machinery health prognostics: A systematic review from data acquisition to rul prediction, *Mechanical systems and signal processing* 104 (2018) 799–834.
- [4] J. Park, D. Lee, Single image-based reflection removal via dual-stream multi-column reversible encoding, *Applied Sciences* 15 (2025) 11229.

- [5] A. Saxena, K. Goebel, D. Simon, N. Eklund, Damage propagation modeling for aircraft engine run-to-failure simulation, in: 2008 international conference on prognostics and health management, IEEE, 2008, pp. 1–9.
- [6] G. Sateesh Babu, P. Zhao, X.-L. Li, Deep convolutional neural network based regression approach for estimation of remaining useful life, in: International conference on database systems for advanced applications, Springer, 2016, pp. 214–228.
- [7] S. Zheng, K. Ristovski, A. Farahat, C. Gupta, Long short-term memory network for remaining useful life estimation, in: 2017 IEEE international conference on prognostics and health management (ICPHM), IEEE, 2017, pp. 88–95.
- [8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, *Information fusion* 58 (2020) 82–115.
- [9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* 51 (2018) 1–42.
- [10] L. Aversano, I. Mancino, A. Marengo, C. Verdone, Comparative analysis of machine learning and deep learning models for atrial fibrillation detection from long-term ecg, *Applied Sciences* 16 (2026) 2390.
- [11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* 30 (2017).
- [12] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.
- [13] M. Raissi, P. Perdikaris, 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* 378 (2019) 686–707.
- [14] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), volume 8, 2017, pp. 8–13.
- [15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, *arXiv preprint arXiv:2108.07258* (2021).
- [16] A. Giuliano, A. McCafferty-Leroux, J. Yawney, S. A. Gadsden, Cognitive internet of things: A review of theory, applications, and recent advances, *IEEE Communications Surveys & Tutorials* (2025).
- [17] W. Wu, J. Hu, J. Zhang, Prognostics of machine health condition using an improved arima-based prediction method, in: 2007 2nd IEEE Conference on Industrial Electronics and Applications, Ieee, 2007, pp. 1062–1067.
- [18] T. H. Loutas, D. Roulias, G. Georgoulas, Remaining useful life estimation in rolling bearings utilizing data-driven probabilistic e-support vectors regression, *IEEE Transactions on Reliability* 62 (2013) 821–832.
- [19] S. Zhao, Y. Zhang, S. Wang, B. Zhou, C. Cheng, A recurrent neural network approach for remaining useful life prediction utilizing a novel trend features construction method, *Measurement* 146 (2019) 279–288.
- [20] Y. Liu, C. Li, M. Li, Data-driven online state prediction method for the traction motors of electric multiple units (emus), *Sustainability* 17 (2025) 4237.
- [21] Z. Chen, M. Wu, R. Zhao, F. Guretno, R. Yan, X. Li, Machine remaining useful life prediction via an attention-based deep learning approach, *IEEE Transactions on Industrial Electronics* 68 (2020) 2521–2531.
- [22] A. Wahid, M. Yahya, J. G. Breslin, M. A. Intizar, Self-attention transformer-based architecture for remaining useful life estimation of complex machines, *Procedia Computer Science* 217 (2023) 456–464.
- [23] R. Avsar, Enhancing prognostic model interpretability for advanced engine failure prediction using prognostic metrics and explainable ai, *The Aeronautical Journal* (2026) 1–22.

- [24] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
- [25] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, arXiv preprint arXiv:1708.08296 (2017).
- [26] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature machine intelligence* 1 (2019) 206–215.
- [27] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.
- [28] A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, arXiv preprint arXiv:1703.03717 (2017).
- [29] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, *Nature machine intelligence* 3 (2021) 620–631.
- [30] E. Weinberger, J. Janizek, S.-I. Lee, Learning deep attribution priors based on prior knowledge, *Advances in Neural Information Processing Systems* 33 (2020) 14034–14045.
- [31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [32] X. Xie, J. Wang, Y. Han, W. Li, Knowledge graph-based in-context learning for advanced fault diagnosis in sensor networks, *Sensors* 24 (2024) 8086.
- [33] Y. Li, H. Zhao, H. Jiang, Y. Pan, Z. Liu, Z. Wu, P. Shu, J. Tian, T. Yang, S. Xu, et al., Large language models for manufacturing, arXiv preprint arXiv:2410.21418 (2024).
- [34] W. Chen, J. Zhang, H. Liu, M. Wang, Y. Li, Can large language models diagnose machine faults? a comprehensive survey from deep learning to foundation models (2026).
- [35] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM computing surveys* 55 (2023) 1–38.
- [36] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780.
- [37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
- [38] Z. Cheng, H. Xiang, P. Ma, L. Zeng, X. Jin, X. Yang, J. Lin, X. Feng, Y. Deng, C. Deng, et al., Maskmol: knowledge-guided molecular image pre-training framework for activity cliffs with pixel masking, *BMC biology* 23 (2025) 279.

A. Appendix

For reproducibility, this appendix provides the mathematical expansion of the SHAP-Informed Dynamic Loss (SIDL) and the pseudo-code for the physics-aware training loop.

A.1. SIDL Gradient Formulation

This section presents the theoretical derivation of the SIDL framework, tracing its formulation from three foundational machine learning paradigms: Cost-Sensitive Learning, Physics-Informed Neural Networks (PINNs), and Explanation Regularization, to its specific application in aerospace prognostics.

Standard neural network training is based on Empirical Risk Minimization (ERM), where the objective is to identify parameters θ that minimize the Mean Squared Error (MSE) across N samples:

$$\mathcal{L}_{MSE}(\theta) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

In standard ERM, all samples receive equal weighting. Inspired by Cost-Sensitive Learning and Focal Loss [37, 38], a dynamic scaling factor $\Omega(\mathbf{x}, \theta)$ is introduced to prioritize specific samples based on their informational value. To define this factor in physical terms, Physics-Informed Neural Networks (PINNs) [13] are considered, which constrain the neural hypothesis space by embedding known Partial Differential Equations (PDEs). Since exact closed-form thermodynamic PDEs are unavailable for the C-MAPSS dataset, the Dual-XAI consensus feature set $\mathcal{C}$ (specifically, sensors s_4 and s_{13}) is used as a data-driven surrogate. The thermodynamic evidence magnitude Φ_i for a given sample is defined as the sum of absolute SHAP values for these consensus sensors:

$$\Phi_i = \sum_{j \in \mathcal{C}} |\phi_{i,j}|$$

Recent advancements in explanation regularization suggest incorporating such interpretability metrics into the loss function [29]. Traditionally, this is accomplished via an additive penalty (e.g., $\mathcal{L}_{add} = \mathcal{L}_{MSE} + \lambda\Phi$). However, an additive penalty creates an adverse optimization landscape for this specific task: to minimize the loss, the optimizer may arbitrarily reduce the network weights associated with the consensus sensors to force $\Phi \rightarrow 0$, which can cause the model to ignore critical physical drivers.

To mathematically circumvent this artifact, the dynamic weight is defined as $\Omega(\mathbf{x}_i) = 1 + \lambda\Phi_i$, resulting in a novel multiplicative formulation:

$$\mathcal{L}_{SIDL} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{MSE}^{(i)} \cdot (1 + \lambda\Phi_i)$$

This formulation guarantees that the XAI penalty remains strictly bounded by the predictive error. As a result, the gradients computed during backpropagation are directly influenced by the thermodynamic evidence:

$$\nabla_{\theta} \mathcal{L}_{SIDL} = (\nabla_{\theta} \mathcal{L}_{MSE}) \cdot \Omega(\mathbf{x}) + \mathcal{L}_{MSE} \cdot (\nabla_{\theta} \Omega(\mathbf{x}))$$

If the model prediction is highly accurate ($\mathcal{L}_{MSE} \rightarrow 0$), the penalty term becomes negligible regardless of the SHAP magnitude. The resulting gradient magnitude is increased by the cognitive scaling factor λ only when the network exhibits a significant predictive error in the presence of pronounced thermodynamic degradation signals. This mechanism compels the LSTM to rapidly abandon spurious statistical minima in favor of physically consistent internal representations.

A.2. Physics-Aware Training Algorithm

Implementation of the SIDL framework requires a custom training loop that dynamically incorporates XAI outputs into the forward pass before gradient computation. The procedure is detailed in Algorithm 1.

Algorithm 1 Physics-Aware SIDL Training Loop

Require: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, Learning rate η , Scaling factor λ , Consensus indices $\mathcal{C}$

Ensure: Optimized Physics-Aware LSTM weights θ^*

```
1: Initialize LSTM weights  $\theta$ 
2: for epoch = 1 to  $E$  do
3:   for each batch  $(\mathbf{x}_{batch}, y_{batch}) \in \mathcal{D}$  do
4:     // Forward Pass
5:      $\hat{y}_{batch} \leftarrow \text{LSTM}(\mathbf{x}_{batch}; \theta)$ 
6:      $\mathcal{L}_{MSE} \leftarrow \frac{1}{B} \sum (y_{batch} - \hat{y}_{batch})^2$ 
7:     // Extract Consensus Physical Impact
8:      $\text{Impact}_{\mathcal{C}} \leftarrow \frac{1}{B} \sum |\mathbf{x}_{batch}[:, :, \mathcal{C}]|$ 
9:     // Compute SIDL
10:     $\mathcal{L}_{SIDL} \leftarrow \mathcal{L}_{MSE} \times (1 + \lambda \times \text{Impact}_{\mathcal{C}})$ 
11:    // Physics-Aware Backpropagation
12:     $g \leftarrow \nabla_{\theta} \mathcal{L}_{SIDL}$ 
13:     $\theta \leftarrow \theta - \eta \cdot g$ 
14:   end for
15: end for
16: return  $\theta^*$ 
```
