Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083272

ARTICLE

PRIME: A Physics-Guided Residual Integrated Framework for Multi-Task Aircraft Engine Diagnostics

Ouail Mjahed[1,*] **and Soukaina Mjahed**[2]

[1]Faculty of Sciences and Technology, Department of Computer Sciences, L2IS Laboratory, Cadi Ayyad University, Marrakech, Morocco

[2]Faculty of Sciences Semlalia, Department of Computer Sciences, LISI Laboratory, Cadi Ayyad University, Marrakech, Morocco

*Corresponding Author: Ouail Mjahed. Email: ouail.mjahed@ced.uca.ma

Received: 31 March 2026; Accepted: 14 May 2026; Published: 30 June 2026

ABSTRACT: Accurate aircraft engine diagnostics is essential for ensuring operational safety and enabling predictive maintenance under heterogeneous operating conditions. Although deep learning models can effectively capture high-dimensional multivariate sensor dynamics, purely data-driven approaches often entangle operating-condition variability with degradation-sensitive patterns, which limits robustness and generalization. This paper introduces **PRIME**, a physics-guided residual integrated framework for multi-task aircraft engine diagnostics. Rather than embedding explicit thermodynamic equations or physical constraints into the optimization process, PRIME relies on a physically motivated residual decomposition strategy that separates operating-condition-driven nominal behavior from degradation-sensitive sensor deviations. Specifically, nominal responses are estimated from operating-condition representations and subtracted from observed sensor signals to isolate fault-relevant residual patterns. These residual representations are then processed by a hybrid temporal architecture combining temporal convolutional networks and transformer-based self-attention, enabling joint modeling of local degradation signatures and long-range temporal dependencies. Within a unified optimization framework, PRIME simultaneously performs Fault Detection (FD), Fault Type Classification (FTC), and Health State Estimation (HSE). Extensive experiments on NASA C-MAPSS, N-CMAPSS, and the ALFA dataset show consistent and statistically significant improvements over strong baseline models. For FD and FTC, PRIME achieves gains of approximately 2%–4% over the strongest neural baselines evaluated under the same protocol, with larger margins over classical machine learning approaches. For HSE, PRIME yields more faithful degradation trajectories, leading to systematic reductions in estimation error across single- and multi-regime datasets. When Remaining Useful Life (RUL) is projected from the learned health trajectory through a threshold-based mechanism, the resulting estimates also improve substantially, with RMSE reductions of up to about 22% under complex operating conditions. These results show that physics-guided residual disentanglement improves robustness, interpretability, and multi-task diagnostic performance. More broadly, they support the view that HSE provides a useful latent degradation representation for downstream prognostic assessment, even though RUL is not directly optimized by the model.

KEYWORDS: Physics-guided learning; residual modeling; multi-task learning; aircraft engine diagnostics; prognostics and health management

1 Introduction

Aircraft engines constitute one of the most safety-critical subsystems in modern aviation, where unexpected failures can severely impact operational safety, mission reliability, and maintenance costs.

To mitigate such risks, Prognostics and Health Management (PHM) systems have become essential for enabling condition-based maintenance and early fault detection. The increasing deployment of onboard sensing technologies has led to the availability of high-dimensional multivariate time-series data describing engine behavior across diverse operating regimes.

Benchmark datasets such as the NASA C-MAPSS turbofan simulation dataset [1] have played a central role in advancing data-driven diagnostic research by providing multivariate run-to-failure trajectories under multiple fault conditions. More recently, the N-CMAPSS dataset [2] introduced higher-fidelity flight condition simulations and richer labeling structures, enabling both prognostics and fault classification under realistic operational variability. In parallel, complementary real-world datasets such as the AirLab Failure and Anomaly (ALFA) dataset [3] provide annotated UAV fault and anomaly scenarios, allowing additional evaluation under practical flight conditions.

Early approaches to aircraft engine fault diagnosis relied primarily on classical machine learning methods, including Support Vector Machines, k-Nearest Neighbors, and ensemble learning techniques [4,5]. While effective under controlled conditions, these models depend heavily on handcrafted features and often struggle to generalize across complex and variable operating regimes.

The emergence of deep learning has significantly improved diagnostic performance by enabling automatic feature extraction from raw multivariate signals. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated strong capabilities in modeling temporal degradation patterns [6,7]. More recently, transformer-based architectures leveraging self-attention mechanisms have shown superior ability to capture long-range temporal dependencies in time-series PHM applications [8,9]. Despite their success, purely data-driven deep models often suffer from limited robustness under varying operating conditions, reduced interpretability, and degraded performance when encountering rare or previously unseen fault modes.

A key challenge in aircraft engine diagnostics lies in the strong coupling between operating-condition dynamics and degradation-induced sensor variations. Changes in altitude, throttle settings, and environmental conditions significantly influence sensor measurements, potentially masking early degradation signatures. Without explicitly accounting for this coupling, data-driven models may inadvertently learn operating regime characteristics rather than fault-related patterns, leading to unreliable generalization.

To address these limitations, hybrid and domain-guided learning paradigms have attracted increasing attention. Theory-guided data science frameworks [10] and physics-informed neural networks [11,12] advocate integrating domain knowledge into data-driven architectures to enhance robustness and interpretability. In the context of aero-engine diagnostics, hybrid approaches combining physics-based performance models with deep learning have demonstrated improved stability under variable operating conditions [13,14]. However, many existing approaches either impose loosely coupled physical constraints or do not explicitly disentangle operating-condition effects from degradation dynamics within the learned representations.

Motivated by the need for robust and generalizable aircraft engine diagnostics under varying operating conditions, this paper proposes **PRIME** (Physics-guided Residual Integrated framework for Multi-task Engine diagnostics), a hybrid framework that explicitly decouples operating-condition dynamics from degradation-sensitive sensor patterns through a *physically motivated residual decomposition strategy*. Specifically, PRIME estimates a nominal sensor response conditioned on the operating regime and subtracts it from the observed measurements to isolate degradation-relevant residual signals prior to feature extraction. These residual representations are subsequently processed using a hybrid temporal architecture combining Temporal Convolutional Networks and transformer-based self-attention, enabling the model to capture both local degradation signatures and long-range temporal dependencies. The proposed nominal–residual

separation is intended to reduce the entanglement between operating-condition variability and degradation-sensitive patterns. However, this decomposition remains an approximation and may become less accurate when strong nonlinear interactions exist between regime dynamics and fault evolution. More broadly, while the framework is evaluated on heterogeneous datasets, the present study does not yet constitute a dedicated stress test for unseen operating regimes or sensor drift. Accordingly, the value of PRIME should be understood not only in terms of average accuracy gains, but also in its unified treatment of FD, FTC, and HSE, its support for downstream prognostic analysis, and its built-in interpretability under variable operating conditions. In addition, a hierarchical attention mechanism provides sensor-level and temporal interpretability, improving diagnostic transparency and trustworthiness.

Unlike conventional regime normalization techniques, which attempt to mitigate operating-condition variability through preprocessing transformations or domain adaptation strategies, PRIME performs representation-level disentanglement through residual modeling. By explicitly separating nominal regime-dependent behavior from degradation-induced variations, the framework allows the neural network to focus more directly on fault-related information.

It is important to distinguish the proposed approach from physics-informed neural network (PINN) paradigms in the strict sense. PINN-based methods typically embed governing equations, conservation laws, or explicit physical constraints into the optimization process. PRIME does not require explicit thermodynamic equations or equation-constrained loss terms. Instead, its physical grounding lies in a structured and physically motivated separation between operating-condition-driven nominal behavior and degradation-sensitive residual deviations. In this sense, PRIME is best characterized as a *physics-guided residual learning framework* rather than a physics-informed model in the strict PINN sense.

Consequently, PRIME combines the flexibility of data-driven deep learning with a physically motivated residual representation, enabling improved robustness under varying operating conditions while remaining compatible with practical PHM settings in which explicit first-principles models may be unavailable or difficult to integrate.

The main contributions of this work are summarized as follows:

- We propose a physics-guided residual decomposition strategy that explicitly separates operating-condition effects from degradation-sensitive sensor variations in aircraft engine data.
- We design a hybrid deep architecture integrating temporal convolution and transformer-based self-attention for multi-scale temporal modeling.
- We incorporate a hierarchical attention mechanism to enhance interpretability and robustness of diagnostic decisions.
- We develop a unified multi-task framework capable of simultaneously performing fault detection (FD), fault type classification (FTC), and health state estimation (HSE) within a single model.
- We conduct extensive experiments on C-MAPSS, N-CMAPSS, and ALFA datasets, demonstrating consistent and statistically significant improvements over strong machine learning and deep learning baselines under a unified evaluation protocol.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed PRIME framework. Section 4 describes the experimental setup, while Section 5 presents and discusses the experimental results. Sections 6 and 7 report the statistical significance analysis and the interpretability/physical consistency analysis, respectively. Section 8 presents the ablation study, Section 9 analyzes the computational complexity, Section 10 discusses the comparison with related literature, Section 11 outlines the main limitations of the proposed approach, and Section 12 concludes the paper.

2 Related Work

This section reviews recent advances in aircraft engine fault diagnosis and prognostics with a focus on five complementary directions: deep temporal modeling, transformer-based PHM architectures, robustness under varying operating conditions, residual and physics-guided hybrid learning, and explainability in safety-critical PHM systems.

2.1 Deep Temporal Modeling for Engine Fault Diagnosis

Deep temporal models have significantly advanced aircraft engine fault diagnosis by enabling direct learning from multivariate sensor trajectories. Beyond early CNN- and LSTM-based architectures, recent studies have explored more expressive temporal modeling strategies for capturing progressive degradation patterns. Multi-scale convolutional networks and dilated Temporal Convolutional Networks (TCNs) have shown improved capability in capturing local and medium-range degradation dynamics at different temporal resolutions [15,16].

Graph-based deep learning approaches have also been introduced to explicitly model inter-sensor dependencies. In particular, Graph Neural Networks (GNNs) have shown promising results in engine fault classification by learning relational structures among correlated sensor channels [17,18]. These developments highlight the importance of jointly modeling temporal evolution and cross-sensor interactions.

However, most temporal models remain predominantly data-driven and often assume stationary degradation structures. As a result, they may not explicitly account for operating-condition variability, which can lead to entanglement between regime effects and degradation-sensitive patterns.

2.2 Transformer-Based Architectures in PHM

Transformer models have recently gained increasing attention in PHM due to their ability to capture long-range dependencies without recurrent computations. Architectures such as Informer, Autoformer, and related time-series transformers have been adapted to Remaining Useful Life (RUL) prediction and health state estimation [19,20].

In engine diagnostics, attention-based architectures can improve feature weighting across sensors and time steps, while also providing a degree of interpretability by highlighting critical degradation windows [21]. These properties make transformers attractive for complex PHM settings in which both short-term fluctuations and long-horizon dependencies are important.

Recent work has also explored the use of time-series foundation models for few-shot aircraft engine prognostics, highlighting the growing importance of transferability and data-efficient temporal representation learning in PHM settings [22]. Current aircraft engine PHM studies have explored adaptive deep Q-learning and heterogeneous deep ensembles, further underscoring the need for operating-condition-aware and unified diagnostic frameworks [23,24].

Nevertheless, transformer-based models remain largely data-driven and typically exhibit quadratic complexity with sequence length. More importantly, most existing transformer PHM studies focus on prognostics, especially RUL estimation, rather than explicit multi-class fault diagnosis under heterogeneous operating conditions. Their robustness under strong regime shifts also remains an open challenge.

2.3 Robustness under Variable Operating Conditions

A persistent difficulty in aircraft engine diagnostics is the strong influence of operating conditions on sensor measurements. Variations in altitude, throttle setting, ambient pressure, and flight regime can induce substantial changes in the observed signals, often masking early degradation signatures.

To address this issue, several normalization and domain adaptation strategies have been proposed. Domain-adversarial neural networks and transfer learning approaches have been applied to C-MAPSS subsets with multiple operating regimes to reduce distribution shifts between training and testing data [25]. Other studies employ operating-condition clustering or regime-specific normalization prior to model training [26].

Although these strategies can mitigate regime bias, they often treat operating conditions as nuisance variables to be normalized away, rather than explicitly separating nominal operating behavior from degradation-related deviations. Consequently, degradation-sensitive information may remain entangled with regime-dependent variability in the learned latent space.

2.4 Residual Learning and Physics-Guided Hybrid Frameworks

Residual-based modeling has long been recognized as an effective strategy for isolating fault-relevant deviations from nominal behavior. In aero-engine monitoring, residual generation techniques derived from performance analysis or reference models have been used for fault isolation in model-based diagnosis frameworks [27].

More recently, hybrid approaches combining physical insight with neural networks have been proposed to improve robustness and interpretability. For example, physics-augmented deep models incorporate estimated thermodynamic variables or simulator-derived features as additional inputs to improve fault classification performance [13]. Other approaches use physics-aware regularization to encourage consistency between model outputs and known physical relationships [28,29].

It is important, however, to distinguish between *physics-informed learning* in the strict PINN sense and broader *physics-guided* or *physically motivated* hybrid modeling. PINN-style methods explicitly embed governing equations, conservation laws, or differentiable physical constraints into the optimization process [11,12]. By contrast, many practical PHM methods do not impose explicit equations, but instead exploit domain knowledge through architecture design, feature engineering, or residual decomposition. PRIME belongs to this latter category: its physical motivation lies in separating operating-condition-driven nominal behavior from degradation-sensitive residual deviations, rather than enforcing thermodynamic equations in the loss function.

Despite these advances, limited work has systematically combined residual-based operating-condition disentanglement with hybrid temporal modeling and built-in interpretability for multi-task aircraft engine diagnostics.

2.5 Explainability in PHM Models

As deep PHM models become increasingly complex, interpretability has become essential, particularly in safety-critical applications. Attention visualization, SHAP values, saliency analysis, and related post-hoc techniques have been used to identify important sensors and degradation stages in engine diagnostics [21,30]. However, post-hoc explanation methods do not always guarantee faithful correspondence with the internal reasoning process of the model. This limitation has motivated interest in architectures that incorporate interpretability directly into the learning pipeline. In aircraft engine diagnostics, such intrinsically interpretable designs remain relatively underexplored.

2.6 Positioning of PRIME

In contrast to prior work, **PRIME** introduces a physics-guided residual decomposition mechanism that explicitly separates operating-condition-driven nominal behavior from degradation-sensitive sensor deviations prior to deep temporal feature extraction. Unlike preprocessing-based regime normalization or

domain adaptation methods, PRIME modifies the representation space itself through nominal–residual separation, allowing the downstream network to focus more directly on fault-relevant information.

Furthermore, PRIME integrates this residual modeling strategy within a hybrid Temporal Convolutional–Transformer architecture equipped with hierarchical attention. This combination enables:

- multi-scale temporal degradation modeling across short- and long-range dependencies,
- improved robustness under heterogeneous operating regimes,
- simultaneous multi-task predictions for FD, FTC, and HSE,
- built-in interpretability at both sensor and temporal levels.

Accordingly, the contribution of PRIME is not to introduce physics-informed learning in the strict PINN sense, but to provide a unified physics-guided residual learning framework for interpretable and robust multi-task aircraft engine diagnostics.

3 Proposed Methodology

This section presents the proposed **PRIME** framework (*Physics-guided Residual Integrated Multi-task framework for aircraft Engine diagnostics*), a physics-guided and domain-robust deep learning framework designed for aircraft engine diagnostics under heterogeneous operating conditions.

PRIME is built around four main design principles: (i) explicit disentanglement of operating-condition effects from degradation-sensitive sensor variations, (ii) physically motivated residual decomposition, (iii) hybrid temporal modeling of residual dynamics, and (iv) unified multi-task prediction of fault detection (FD), fault type classification (FTC), and health state estimation (HSE).

Fig. 1 illustrates the overall workflow of PRIME. At a high level, the framework first estimates nominal operating-condition-driven behavior, then isolates degradation-sensitive residuals, and finally processes these residual representations through a hybrid temporal encoder with hierarchical attention to support multi-task diagnostic inference.

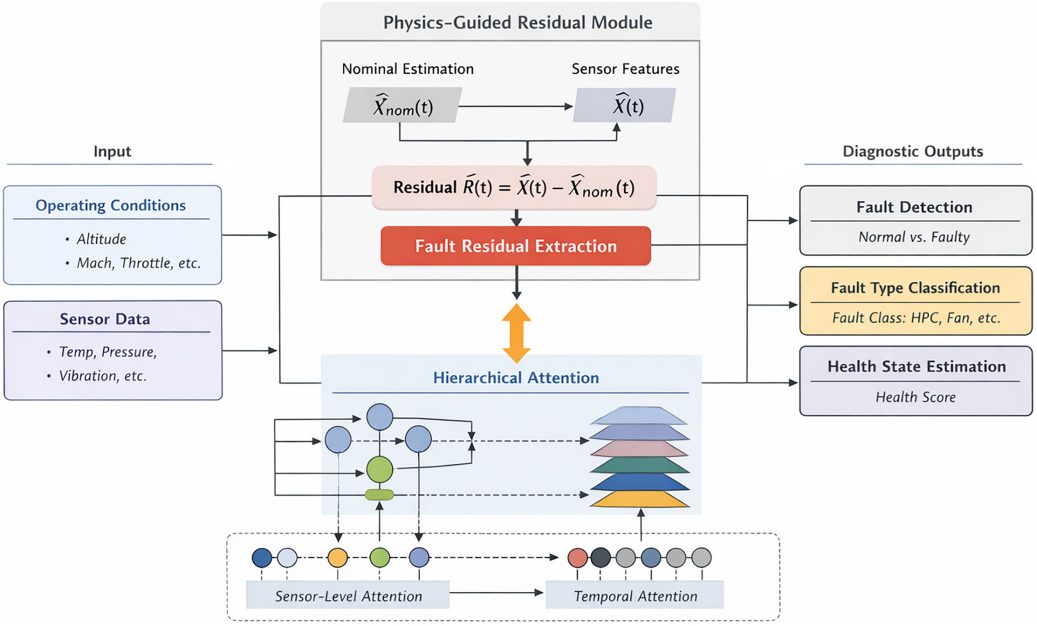

Figure 1: PRIME framework: physics-guided residual decomposition with hierarchical attention for multi-task aircraft engine diagnostics.

3.1 Problem Formulation

Let $X = \{x_1, x_2, \ldots, x_T\}$ denote a multivariate time-series of engine measurements, where each observation $x_t \in \mathbb{R}^S$ contains S sensor signals at time step t. The diagnostic objective is to infer the engine health state y_t from historical observations:

$$y_t = f(X_{1:t}), \tag{1}$$

where $f(\cdot)$ denotes a nonlinear mapping learned from data. To explicitly account for operating-condition variability, each observation is decomposed as:

$$x_t = [o_t, s_t], \tag{2}$$

where o_t denotes operating-condition variables (e.g., altitude, Mach number, throttle setting), and s_t represents the thermodynamic and mechanical sensor measurements.

3.2 Neural Architecture of PRIME

PRIME is organized as a two-branch neural architecture followed by a shared latent representation and three task-specific prediction heads. The first branch encodes operating-condition variables in order to estimate nominal behavior under varying regimes. The second branch processes degradation-sensitive residual signals through a hybrid temporal encoder that combines local and long-range sequence modeling.

The attended shared representation is then exploited by three heads dedicated to FD, FTC, and HSE. Importantly, PRIME does *not* directly optimize RUL through an additional neural head. Instead, RUL is treated as a downstream prognostic quantity projected from the learned health trajectory, as described in Section 3.8. Table 1 summarizes the functional neural architecture of the framework.

Table 1: Functional neural architecture of the proposed PRIME framework.

Module	Input	Output	Function
Operating-condition encoder	o_t	$h_t^{(o)}$	Encodes regime-dependent nominal operating dynamics
Nominal projection layer	$h_t^{(o)}$	$\hat{s}_t^{nominal}$	Estimates nominal sensor response
Residual decomposition	$s_t, \hat{s}_t^{nominal}$	r_t	Isolates degradation-sensitive deviations
Residual temporal encoder	$r_{1:t}$	$h_t^{(s)}$	Models local and long-range fault dynamics
Sensor-level attention	$h_t^{(s)}$	α_i	Assigns sensor importance weights
Temporal attention	$h_t^{(s)}$	β_t	Highlights critical degradation stages
Shared latent representation	attended features	z	Provides unified representation for downstream tasks
FD head	z	$\hat{y}_{FD}$	Predicts healthy vs. faulty condition
FTC head	z	$\hat{y}_{FTC}$	Predicts fault category
HSE head	z	$\hat{y}_{HSE}$	Estimates continuous health state
RUL projection (downstream)	$\hat{H}(t)$	$\widehat{RUL}(t)$	Projects RUL from the learned health trajectory

Because PRIME follows a sequential pipeline in which operating-condition encoding precedes residual extraction and temporal modeling, errors introduced in upstream stages may influence downstream

representations. This modular design was adopted to improve interpretability and explicit nominal–residual disentanglement, but it also implies that imperfect nominal estimation may affect subsequent diagnostic processing.

3.3 Operating-Condition Encoding

The operating-condition branch captures regime-dependent nominal dynamics driven by environmental and control variables. A gated recurrent unit (GRU) is used to model temporal dependencies in the operating-condition sequence [31,32]:

$$h_t^{(o)} = \mathrm{GRU}(o_t, h_{t-1}^{(o)}), \tag{3}$$

where $h_t^{(o)}$ denotes the latent encoding of nominal operating behavior. This representation is not intended to capture degradation directly; rather, it serves as a basis for estimating the expected sensor response under the current operating regime.

3.4 Physics-Guided Residual Decomposition

To isolate degradation-sensitive deviations from operating-condition-driven variability, PRIME introduces a physics-guided residual decomposition layer. The approach is *physically motivated* rather than equation-constrained: it does not embed explicit thermodynamic equations or conservation laws in the loss function. Instead, it relies on the assumption that observed sensor measurements can be decomposed into: (i) a nominal component primarily associated with operating conditions, and (ii) a residual component more sensitive to degradation.

*Nominal sensor behavior*is estimated as:

$$\hat{s}_t^{nominal} = g\left(h_t^{(o)}\right), \tag{4}$$

where $g(\cdot)$ is a learnable projection from the operating-condition latent representation.

The residual signal is then computed as:

$$r_t = s_t - \hat{s}_t^{nominal}. \tag{5}$$

This residual decomposition relies on an approximate additive separation between regime-dependent nominal behavior and degradation-sensitive deviations. Such an assumption is physically motivated and practically useful for disentangling operating-condition effects from fault-related patterns, but it does not imply that engine dynamics are strictly additive in all regimes. In particular, strong nonlinear coupling between operating conditions and degradation mechanisms may reduce the fidelity of this decomposition.

In addition, the quality of the residual representation depends on the accuracy of the nominal estimation stage. If the operating-condition encoder does not fully capture regime-dependent behavior, the subtraction step may introduce residual bias, which can then propagate to the downstream temporal modeling and task-specific heads. Accordingly, the proposed residual representation should be interpreted as a structured approximation for improving diagnostic inference rather than as an exact physical decomposition of engine behavior.

This decomposition suppresses operating-condition effects and highlights fault-relevant deviations. As a result, the downstream temporal encoder receives inputs that are less dominated by regime variability and more directly linked to degradation-sensitive patterns.

3.5 Residual Temporal Encoder

The residual sequence is processed through a hybrid temporal encoder composed of a Temporal Convolutional Network (TCN) followed by a Transformer encoder. The TCN captures short-range and local temporal correlations [15,33], while the Transformer models long-range dependencies across time and residual patterns [34]:

$$h_t^{(s)} = \text{Transformer}\big(\text{TCN}(r_{1:t})\big). \tag{6}$$

This hybrid design combines the strengths of convolutional temporal abstraction and self-attention-based sequence modeling, enabling more effective representation of progressive degradation signatures.

3.6 Hierarchical Attention Mechanism

To enhance interpretability and diagnostic precision, PRIME applies hierarchical attention over the encoded residual features [34,35]. Two complementary attention mechanisms are used.

Sensor-level attention assigns adaptive importance weights to each sensor channel:

$$\alpha_i = \frac{\exp(w^\top r_i)}{\sum_{j=1}^{S} \exp(w^\top r_j)}, \tag{7}$$

where α_i measures the relative contribution of the i-th sensor to the diagnostic decision.

Temporal attention identifies critical time steps associated with degradation progression:

$$\beta_t = \frac{\exp(u^\top r_t)}{\sum_{\tau=1}^{T} \exp(u^\top r_\tau)}. \tag{8}$$

The combination of sensor-level and temporal attention yields an attended representation that is both more discriminative and more interpretable.

3.7 Multi-Task Diagnostic Framework

PRIME is formulated as a unified multi-task learning framework addressing three complementary objectives: **Fault Detection (FD)**, **Fault Type Classification (FTC)**, and **Health State Estimation (HSE)**.

Given an input sequence $X \in \mathbb{R}^{T\times d}$, the shared encoder produces a latent representation:

$$z = f_{\text{PRIME}}(X) \in \mathbb{R}^h, \tag{9}$$

which feeds three task-specific heads:

$$\hat{y}_{FD} = \sigma(W_{FD} z + b_{FD}), \tag{10}$$

$$\hat{y}_{FTC} = \text{Softmax}(W_{FTC} z + b_{FTC}), \tag{11}$$

$$\hat{y}_{HSE} = \phi(W_{HSE} z + b_{HSE}), \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid activation and $\phi(\cdot)$ is a bounded activation ensuring $\hat{y}_{HSE} \in [0,1]$.

HSE as a Continuous Diagnostic Variable: HSE provides a continuous representation of degradation dynamics. Rather than treating health as a direct rescaling of RUL, PRIME learns a relative health index reflecting the progressive deviation of the engine from nominal behavior. Healthy conditions correspond to values close to 1, while increasingly degraded states approach 0. In multi-fault settings, intermediate values may capture relative degradation severity.

Because the HSE head is trained from residual degradation-sensitive representations, it is expected to encode smoother degradation dynamics than purely class-based supervision alone. In the main evaluation protocol, HSE is treated primarily as a continuous regression task. Accordingly, the predicted health index $\hat{y}_{HSE} \in [0,1]$ is evaluated using RMSE, MAE, and R^2, which quantify trajectory fidelity and estimation accuracy. Threshold-based discretization of HSE, when used, is considered only as a supplementary analysis and is explicitly defined in the corresponding experimental setting.

Joint Optimization Objective: The overall training objective combines the three task-specific losses:

$$\mathcal{L}_{total} = \lambda_{FD}\mathcal{L}_{FD} + \lambda_{FTC}\mathcal{L}_{FTC} + \lambda_{HSE}\mathcal{L}_{HSE}, \tag{13}$$

where $\mathcal{L}_{FD}$ is Binary Cross-Entropy, $\mathcal{L}_{FTC}$ is Categorical Cross-Entropy, and $\mathcal{L}_{HSE}$ is Mean Squared Error. The coefficients λ_{FD}, λ_{FTC}, and λ_{HSE} balance the three objectives during joint optimization.

3.8 From Health State Estimation to RUL Projection

PRIME does not directly optimize Remaining Useful Life (RUL). Instead, RUL is estimated as a downstream prognostic projection from the learned health trajectory $\hat{H}(t) \in [0,1]$, where higher values indicate healthier operating conditions and lower values indicate progressive degradation.

Failure threshold: A failure event is defined as the first time at which the predicted health trajectory reaches a predefined threshold $\tau \in (0,1)$. In this work, τ is treated as a calibration parameter selected on the validation set and then fixed for all test experiments. Formally, the predicted failure time is defined as:

$$\hat{t}_{fail} = \inf\left\{t' > t \mid \hat{H}(t') \le \tau\right\}. \tag{14}$$

The estimated RUL at time t is then computed as:

$$\widehat{RUL}(t) = \hat{t}_{fail} - t. \tag{15}$$

Monotonic extrapolation: In practical settings, the predicted health trajectory may not intersect the threshold within the observed time horizon. In such cases, $\hat{t}_{fail}$ is estimated using a local monotonic extrapolation fitted over the most recent degradation window of length w. Specifically, a monotonic linear approximation is fitted to the last w points of $\hat{H}(t)$, and the intersection between the extrapolated trajectory and the threshold τ is used to estimate $\hat{t}_{fail}$.

Let the fitted local degradation model be:

$$\tilde{H}(t) = at + b, \tag{16}$$

with $a < 0$ enforced to preserve monotonic degradation. The extrapolated failure time is then obtained as:

$$\hat{t}_{fail} = \frac{\tau - b}{a}. \tag{17}$$

Evaluation protocol: RUL performance is evaluated using RMSE, MAE, and the NASA scoring function. Importantly, no additional RUL loss is introduced during training. Therefore, RUL should be interpreted as a downstream operational indicator derived from the learned health trajectory rather than as an independently optimized prediction task.

Discussion. More flexible nonlinear projection strategies could also be considered, such as spline-based extrapolation, piecewise degradation models, Gaussian-process trajectory forecasting, or parametric wear models. Such alternatives may better capture nonlinear degradation phases, especially in settings with strong

regime changes or accelerated wear near failure. Exploring these nonlinear HSE-to-RUL mappings is left for future work.

3.9 Algorithmic Summary

Algorithm 1 summarizes the training and inference procedure of PRIME. The framework first estimates nominal operating-condition-driven behavior, then extracts degradation-sensitive residuals, and finally processes these residual representations through a hybrid temporal encoder with hierarchical attention for multi-task prediction.

Algorithm 1: PRIME training and inference procedure

Require: Multivariate time-series $X = \{(o_t, s_t)\}_{t=1}^{T}$, labels y_{FD}, y_{FTC}, y_{HSE}, learning rate η, task weights $\lambda_{FD}, \lambda_{FTC}, \lambda_{HSE}$

Ensure: Fault detection $\hat{y}_{FD}$, fault type classification $\hat{y}_{FTC}$, health estimation $\hat{y}_{HSE}$

Initialize parameters: θ_o (GRU operating-condition encoder), θ_s (residual temporal encoder), θ_a (hierarchical attention module)
Initialize task heads: $\theta_{FD}, \theta_{FTC}, \theta_{HSE}$
for each training epoch **do**
  for each mini-batch $\mathcal{B}$ **do**
    Encode operating conditions: $h_t^{(o)} \leftarrow GRU(o_t; \theta_o)$
    Estimate nominal sensor behavior: $\hat{s}_t^{nominal} \leftarrow g(h_t^{(o)})$
    Compute residual sequence: $r_t \leftarrow s_t - \hat{s}_t^{nominal}$
    Encode residual temporal dynamics: $h_t^{(r)} \leftarrow Transformer(TCN(r_{1:t}; \theta_r))$
    Compute sensor-level attention: $\alpha_{t,i} \leftarrow Softmax(W_s h_t^{(r)})$
    Compute temporal attention: $\beta_t \leftarrow Softmax(W_t h_t^{(r)})$
    Form attended shared representation: $z \leftarrow \sum_{t=1}^{T} \beta_t \left(\sum_{i=1}^{S} \alpha_{t,i} h_{t,i}^{(r)} \right)$
    Task predictions: $\hat{y}_{FD} \leftarrow Head_{FD}(z)$, $\hat{y}_{FTC} \leftarrow Head_{FTC}(z)$, $\hat{y}_{HSE} \leftarrow Head_{HSE}(z)$
    Compute losses: $\mathcal{L}_{FD} \leftarrow BCE(y_{FD}, \hat{y}_{FD})$, $\mathcal{L}_{FTC} \leftarrow CCE(y_{FTC}, \hat{y}_{FTC})$, $\mathcal{L}_{HSE} \leftarrow MSE(y_{HSE}, \hat{y}_{HSE})$
    Total loss: $\mathcal{L} \leftarrow \lambda_{FD}\mathcal{L}_{FD} + \lambda_{FTC}\mathcal{L}_{FTC} + \lambda_{HSE}\mathcal{L}_{HSE}$
    Update all parameters using Adam optimizer
  end for
end for
Inference: output $(\hat{y}_{FD}, \hat{y}_{FTC}, \hat{y}_{HSE})$

4 Experimental Setup

This section describes the datasets, preprocessing protocol, implementation details, evaluation metrics, and hyperparameter configuration used to validate the proposed PRIME framework.

4.1 Datasets Description

The proposed PRIME framework is evaluated on three benchmark datasets covering both simulated turbofan degradation scenarios and complementary real-world UAV fault conditions: C-MAPSS, N-CMAPSS, and ALFA. Together, these datasets support a broad assessment of diagnostic robustness under controlled simulation settings and practical flight disturbances.

C-MAPSS: The NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset [1,36] provides multivariate run-to-failure time-series collected from simulated turbofan engines under controlled degradation scenarios. The benchmark consists of trajectories generated under three operating settings and 21 sensor measurements, together with unit index and cycle information, resulting in 26 columns in the raw data format. Four subsets (FD001–FD004) are provided, differing in operating conditions and fault modes. FD002 and FD004 involve multiple operating regimes, while FD003 and FD004 include two degradation modes, namely HPC degradation and fan degradation [37].

N-CMAPSS: The N-CMAPSS dataset [2,38] extends C-MAPSS with higher-fidelity physics-based simulation and more realistic flight profiles. It includes additional environmental variables, richer sensor measurements, and explicit fault annotations. Compared with C-MAPSS, N-CMAPSS introduces greater operational variability and more complex degradation behavior, making it particularly relevant for evaluating robustness under realistic flight-condition changes.

ALFA: The AirLab Failure and Anomaly (ALFA) dataset [3] is a real fixed-wing UAV fault and anomaly detection dataset comprising 47 autonomous flight sequences collected under multiple failure scenarios. These include sudden full engine power loss as well as control-surface actuator faults. Unlike the simulation-based C-MAPSS and N-CMAPSS benchmarks, ALFA reflects real-world noise, non-stationarity, and operational uncertainty. In this work, ALFA is used as a complementary real-world dataset for evaluating fault detection robustness under practical flight conditions, rather than as a full turbofan prognostics benchmark.

Table 2 summarizes the key characteristics of the six aircraft engine datasets employed for evaluating the PRIME framework.

Table 2: Summary of datasets and experimental tasks.

Dataset	# Units	Op. Conditions	# Fault Modes	# Sensors	Tasks
FD001 (C-MAPSS)	100	1	1	21	FD, HSE
FD002 (C-MAPSS)	260	6	1	21	FD, HSE
FD003 (C-MAPSS)	100	1	2	21	FD, FTC, HSE
FD004 (C-MAPSS)	249	6	2	21	FD, FTC, HSE
N-CMAPSS	90+	Multiple	Multiple	30+	FD, HSE
ALFA	47 flights	Variable	Multiple	20+	FD

Note: N-CMAPSS is used for FD and HSE in this study. Although the dataset contains richer fault-related annotations, the adopted subset does not provide a fault taxonomy directly aligned with the FTC configuration used for FD003 and FD004. ALFA is used only for FD. Although it provides real-world fault/anomaly annotations from fixed-wing UAV flights, it does not include continuous health labels or standardized run-to-failure trajectories required for a consistent HSE and RUL benchmark.

4.2 Data Preparation and Preprocessing

To ensure methodological consistency across datasets and tasks, all experiments in this work are conducted under a unified ten-fold cross-validation protocol at the engine-unit level. For each fold, engine trajectories are partitioned into training and validation/test subsets without overlap, so that all reported results are expressed as mean ± standard deviation across folds. Each sample includes an engine unit identifier, a time-step index, operational settings, and multivariate sensor measurements that characterize the degradation process. These signals capture essential thermodynamic and mechanical indicators, enabling the modeling of degradation progression from healthy operation to gradual wear and eventual failure.

A unified preprocessing pipeline is then applied across datasets and tasks.

4.2.1 Signal Normalization

Sensor signals are standardized using z-score normalization:

$$\tilde{x} = \frac{x - \mu}{\sigma}, \tag{18}$$

where μ and σ are computed exclusively from the training set to prevent data leakage. For datasets with multiple operating conditions (e.g., FD002 and FD004), normalization is performed condition-wise.

4.2.2 Temporal Segmentation

A sliding-window strategy is applied to construct temporal samples:

$$X_k = x_k, x_{k+1}, \ldots, x_{k+T-1}. \tag{19}$$

This segmentation enables the learning of short- and mid-term degradation dynamics.

4.2.3 Label Construction

Labels are generated according to the task formulation:

- **FD:** binary labels derived from fault annotations or failure-oriented decision criteria, depending on the dataset.
- **FTC:** multi-class labels corresponding to distinct fault modes (FD003, FD004).
- **HSE:** continuous health targets defined as a relative degradation indicator, where values close to 1 correspond to healthy operation and lower values indicate progressive degradation.

For ALFA, noise filtering and outlier removal are additionally applied to mitigate measurement variability inherent to real flight data.

4.3 Implementation Details

The PRIME model is implemented in `PyTorch`. Experiments are conducted on an NVIDIA RTX A6000 GPU (48 GB VRAM).

Model parameters are optimized using the Adam optimizer [39].

Training is performed for a maximum of 100 epochs with early stopping based on validation loss to mitigate overfitting.

A ten-fold cross-validation protocol was adopted at the engine-unit level. For each fold, engines were partitioned into training and validation/test subsets without overlap, and all reported results are expressed as mean ± standard deviation across the ten folds.

4.4 Evaluation Metrics

Performance evaluation is conducted separately for the three diagnostic tasks: FD, FTC, and HSE. Classification metrics are derived from the confusion matrix, where C_{ij} denotes the number of samples belonging to class i predicted as class j.

4.4.1 Classification Metrics (FD, FTC)

For each class i, Precision (P_i), Recall (R_i), and F_1-score ($F_{1,i}$) are defined as:

$$P_i = \frac{C_{ii}}{\sum_j C_{ji}}, \quad R_i = \frac{C_{ii}}{\sum_j C_{ij}}, \quad F_{1,i} = \frac{2P_iR_i}{P_i + R_i}. \tag{20}$$

The overall Accuracy (Acc), macro-averaged Recall (R), and macro-averaged F_1-score (F_1) are computed as:

$$Acc = \frac{\sum_i C_{ii}}{\sum_{i,j} C_{ij}}, \quad R = \frac{1}{K}\sum_{i=1}^{K} R_i, \quad F_1 = \frac{1}{K}\sum_{i=1}^{K} F_{1,i}, \tag{21}$$

where K is the number of classes. Macro-averaging is adopted to ensure balanced evaluation under potential class imbalance, particularly for FTC.

For binary FD, the False Alarm Rate (FAR) is defined as:

$$FAR = \frac{FP}{FP + TN}, \tag{22}$$

where FP and TN denote false positives and true negatives, respectively.

The Area Under the Receiver Operating Characteristic Curve (AUC) is additionally reported to evaluate discriminative capability independently of decision thresholds. For multi-class FTC, the main evaluation metrics are Accuracy, Macro-Recall, Macro-F_1, and Macro-AUC.

4.4.2 Regression Metrics (Continuous HSE and RUL)

For continuous Health State Estimation (HSE), performance is evaluated using regression metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), and the NASA scoring function (S). These metrics quantify the accuracy and fidelity of the predicted degradation trajectory.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \quad R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{23}$$

$$S = \sum_{i=1}^{N} \begin{cases} e^{-\frac{y_i - \hat{y}_i}{13}} - 1, & \text{if} y_i - \hat{y}_i < 0 \\ e^{\frac{y_i - \hat{y}_i}{10}} - 1, & \text{if } y_i - \hat{y}_i \geq 0. \end{cases} \tag{24}$$

4.4.3 Statistical Validation

To validate performance differences between PRIME and baseline models, statistical comparisons are conducted across cross-validation folds. Paired Student's t-test and Wilcoxon signed-rank test [40] are applied:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad Z = \frac{W - \mu_W}{\sigma_W}, \tag{25}$$

where $\bar{d}$ and s_d denote the mean and standard deviation of paired differences across n folds, and W, μ_W, σ_W correspond to the Wilcoxon statistic and its associated parameters.

Statistical significance is established at $p < 0.05$. Effect sizes are additionally reported to assess practical relevance.

4.5 Baseline Methods

PRIME is compared against representative deep learning baselines: CNN-based architectures [41], LSTM networks [42], BiLSTM models [43], and Transformer-based models [34,44].

All baselines are trained under identical preprocessing and evaluation protocols to ensure fair comparison.

4.6 Implementation and Hyperparameter Configuration

The hyperparameters of PRIME were selected through validation-based grid search to ensure fair comparison across datasets and baseline models. In order to improve reproducibility, Table 3 reports the main configuration settings associated with the operating-condition encoder, the residual temporal encoder, the optimization procedure, and the multi-task objective.

Table 3: Implementation and hyperparameter settings of PRIME.

Hyperparameter	Value
Input and temporal configuration	
Sliding window length T	30 timesteps
Operating-condition branch	
GRU hidden units	64
Residual temporal encoder	
TCN filters	64
Transformer layers	2
Transformer attention heads	4
Attention projection dimension	64
Dropout rate	0.3
Optimization settings	
Batch size	256
Learning rate η	1×10^{-4}
Optimizer	Adam
Maximum epochs	100
Multi-task learning	
Loss weights $(\lambda_{FD}, \lambda_{FTC}, \lambda_{HSE})$	$(1.0, 1.0, 1.0)$

5 Multi-Task Results and Discussion

This section evaluates PRIME across the three diagnostic tasks: Fault Detection (FD), Fault Type Classification (FTC), and Health State Estimation (HSE), followed by the derived Remaining Useful Life (RUL) analysis.

Unless otherwise stated, all experiments are conducted using a ten-fold cross-validation protocol at the engine-unit level. The reported results are expressed as mean ± standard deviation across the ten folds.

This protocol is used consistently for performance evaluation and for the fold-wise statistical comparisons reported in the subsequent sections.

5.1 Task–Dataset Alignment

Although PRIME is designed as a unified multi-task framework, each task is evaluated only when the corresponding dataset provides annotations that support a meaningful and methodologically consistent assessment.

Fault Detection (FD) is reported on all datasets, since all of them provide labels that allow discrimination between nominal and faulty operating conditions.

Fault Type Classification (FTC) is evaluated only on FD003 and FD004, which provide a directly comparable multi-class fault taxonomy under controlled simulated degradation settings. Although N-CMAPSS contains richer fault-related annotations, the subset adopted in this study was selected primarily for FD and HSE evaluation under heterogeneous operating conditions. Its label organization is not directly aligned with the FTC configuration used for FD003 and FD004; therefore, N-CMAPSS is not included in the main FTC benchmark in order to preserve task consistency across datasets.

Health State Estimation (HSE) is assessed on FD001–FD004 and N-CMAPSS, where progressive degradation trajectories are available. By contrast, ALFA is not used for HSE or RUL evaluation because it does not provide continuous health trajectories or standardized run-to-failure targets comparable to those available in C-MAPSS and N-CMAPSS.

ALFA is therefore included as a complementary real-world UAV fault/anomaly dataset for evaluating FD robustness under practical flight conditions only. In this sense, PRIME remains a flexible multi-task architecture, but the activation of its task-specific heads depends on the annotation structure and prognostic suitability of each dataset.

This task–dataset alignment ensures methodological consistency, avoids task–label mismatch, and enables fairer interpretation of the reported results. Pairwise statistical significance is further assessed using paired t-tests and Wilcoxon signed-rank tests.

5.2 Fault Detection Performance

Table 4 summarizes the binary FD results across all datasets using ten-fold cross-validation.

Table 4: Ten-fold cross-validation results for fd across all datasets (mean ± std).

Dataset	Method	Acc (%)	R (%)	F_1 (%)	FAR (%)	AUC
FD001	CNN	94.13 ± 0.92	93.87 ± 0.88	93.51 ± 0.84	5.81 ± 0.61	0.962 ± 0.006
	LSTM	94.65 ± 0.81	95.11 ± 0.77	94.92 ± 0.73	4.72 ± 0.54	0.971 ± 0.005
	BiLSTM	95.88 ± 0.74	95.18 ± 0.70	95.63 ± 0.66	4.09 ± 0.49	0.976 ± 0.004
	Transformer	96.81 ± 0.63	96.78 ± 0.60	96.24 ± 0.57	3.78 ± 0.44	0.981 ± 0.004
	PRIME	99.47 ± 0.41	99.37 ± 0.39	99.48 ± 0.36	1.71 ± 0.28	0.993 ± 0.002
FD002	CNN	90.13 ± 1.14	90.62 ± 1.08	90.44 ± 1.03	9.82 ± 0.87	0.921 ± 0.008
	LSTM	92.27 ± 1.02	92.11 ± 0.97	91.92 ± 0.92	8.79 ± 0.79	0.933 ± 0.007
	BiLSTM	94.56 ± 0.95	94.15 ± 0.90	94.58 ± 0.86	8.12 ± 0.71	0.941 ± 0.006
	Transformer	96.13 ± 0.82	96.61 ± 0.78	96.42 ± 0.74	6.89 ± 0.63	0.954 ± 0.005
	PRIME	98.31 ± 0.58	98.82 ± 0.54	98.62 ± 0.51	4.13 ± 0.39	0.982 ± 0.003

(Continued)

Table 4 (continued)

Dataset	Method	Acc (%)	R (%)	F_1 (%)	FAR (%)	AUC
FD003	CNN	92.51 ± 0.88	93.09 ± 0.84	92.89 ± 0.79	6.21 ± 0.58	0.958 ± 0.006
	LSTM	94.23 ± 0.79	93.91 ± 0.75	93.71 ± 0.71	5.62 ± 0.52	0.964 ± 0.005
	BiLSTM	95.10 ± 0.71	94.72 ± 0.67	94.59 ± 0.63	4.83 ± 0.47	0.972 ± 0.004
	Transformer	95.61 ± 0.66	95.21 ± 0.62	95.12 ± 0.59	4.35 ± 0.43	0.978 ± 0.004
	PRIME	98.88 ± 0.47	98.82 ± 0.44	98.79 ± 0.41	2.78 ± 0.31	0.988 ± 0.002
FD004	CNN	89.09 ± 1.21	89.17 ± 1.15	89.12 ± 1.09	11.41 ± 0.96	0.903 ± 0.009
	LSTM	90.31 ± 1.08	90.06 ± 1.02	90.02 ± 0.97	10.22 ± 0.89	0.917 ± 0.008
	BiLSTM	92.33 ± 0.97	92.81 ± 0.92	92.52 ± 0.87	9.09 ± 0.81	0.931 ± 0.007
	Transformer	94.14 ± 0.89	94.49 ± 0.84	94.36 ± 0.79	8.12 ± 0.73	0.942 ±0.006
	PRIME	97.52 ± 0.62	97.68 ± 0.58	97.61 ± 0.54	3.01 ± 0.44	0.975 ± 0.004
N-CMAPSS	CNN	90.31 ± 1.05	90.72 ± 1.00	90.53 ± 0.94	9.81 ± 0.84	0.914 ± 0.008
	LSTM	91.17 ± 0.96	91.14 ± 0.91	91.16 ± 0.86	8.72 ± 0.76	0.938 ± 0.007
	BiLSTM	93.43 ± 0.84	93.48 ± 0.80	93.46 ± 0.75	7.56 ± 0.68	0.949 ± 0.006
	Transformer	95.24 ± 0.72	95.38 ± 0.68	95.36 ± 0.64	6.12 ± 0.56	0.961 ± 0.005
	PRIME	98.42 ± 0.51	98.73 ± 0.48	98.52 ± 0.45	3.79 ± 0.37	0.989 ± 0.003
ALFA	CNN	90.47 ± 1.34	90.69 ± 1.27	90.41 ± 1.21	12.31 ± 1.02	0.907 ± 0.010
	LSTM	91.25 ± 1.21	91.48 ± 1.14	91.37 ± 1.07	11.45 ± 0.94	0.914 ± 0.009
	BiLSTM	93.32 ± 1.02	93.63 ± 0.97	93.49 ± 0.91	10.34 ± 0.86	0.938 ± 0.008
	Transformer	95.51 ± 0.93	95.53 ± 0.88	95.52 ± 0.83	9.56 ± 0.78	0.954 ± 0.007
	PRIME	98.35 ± 0.68	98.29 ± 0.63	98.31 ± 0.59	3.68 ± 0.54	0.984 ± 0.005

As expected, FD001 yields the highest overall performance due to its lower operating-regime variability and simpler degradation patterns compared to multi-condition datasets.

PRIME consistently achieves the highest detection performance across all datasets. Accuracy improvements over the Transformer baseline range from +2.18% (FD002) to +3.38% (FD004). On FD004, PRIME increases the F_1 score from 94.36% to 97.61% while reducing the false alarm rate (FAR) from 8.12% to 3.01%. Similar gains are observed on N-CMAPSS (F_1: 95.36% → 98.52%) and ALFA (95.52% → 98.31%). On average across the six datasets, PRIME improves the F_1 score by approximately 3.1 percentage points over the Transformer baseline while reducing the false alarm rate by more than 50%.

AUC values remain consistently high for PRIME, exceeding 0.98 on five out of six datasets and reaching 0.993 on FD001. Even under the most challenging scenario (FD004), PRIME maintains strong discrimination capability with an AUC of 0.975. Standard deviations remain below 0.7 across all datasets, indicating stable training and low cross-fold variability. Performance gains are particularly pronounced on datasets characterized by higher operating-condition variability (FD003, FD004, and N-CMAPSS), suggesting improved robustness under heterogeneous regimes.

For binary FD, ROC curves are computed for each cross-validation fold. True positive rates (TPR) are interpolated over a common false positive rate (FPR) grid and averaged across folds to obtain robust estimates.

Fig. 2 presents the ROC curves for binary fault detection across three datasets (FD001, FD004, and N-CMAPSS). The curves compare PRIME with baseline models (Transformer, BiLSTM, and CNN).

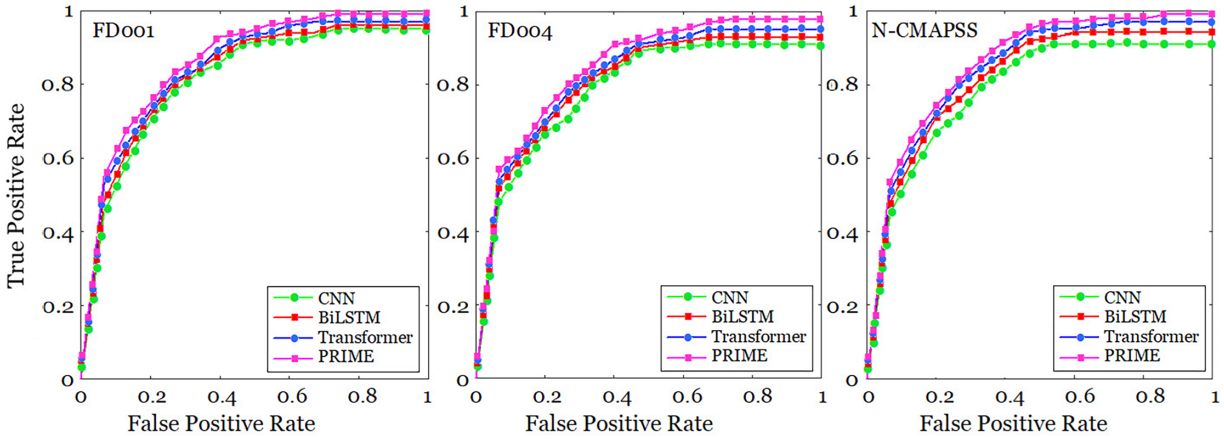

Figure 2: ROC curves for binary FD across FD001, FD004, and N-CMAPSS.

On FD004, PRIME achieves a higher AUC (0.975) compared to the Transformer baseline (0.942), indicating improved discrimination capability. The largest performance gap occurs in the low-FPR region (FPR < 0.1), which is particularly critical for safety-sensitive aerospace applications where false alarms must be minimized.

Although the observed gains over the strongest Transformer baseline remain moderate in absolute terms on some datasets, they are consistent across classification, health estimation, and downstream prognostic indicators. In this sense, the contribution of PRIME should not be interpreted as an accuracy-only improvement, but rather as a more integrated trade-off between multi-task performance, interpretability, and robustness under heterogeneous operating regimes. At the same time, we note that improved Accuracy, F_1, AUC, and MAE are reported as complementary indicators, and their statistical interdependence is not explicitly analyzed in the present study.

5.3 Fault Type Classification Performance

Fault type classification (FTC) is evaluated on FD003 and FD004, which are the two datasets in this study providing a directly comparable multi-class fault taxonomy. N-CMAPSS is not included in the main FTC benchmark because the adopted subset was selected for FD/HSE evaluation and does not align directly with the FTC label structure considered for FD003 and FD004. Table 5 reports the ten-fold cross-validation results.

PRIME consistently achieves the best multi-class classification performance on both FTC benchmarks. On FD003, macro-F_1 increases from 95.35% (Transformer) to 97.89% (+2.54%), while macro-recall rises from 95.57% to 97.98% and accuracy from 95.26% to 97.78%. The macro-AUC also improves from 0.961 to 0.978, indicating enhanced class separability.

The gains are even more pronounced on FD004, where macro-F_1 improves from 93.06% to 96.93% (+3.87%), macro-recall from 93.11% to 96.98%, and accuracy from 92.89% to 96.83%. The macro-AUC correspondingly increases from 0.931 to 0.968. These results confirm that PRIME remains particularly effective under the more challenging multi-regime conditions of FD004.

Taken together, these improvements indicate that PRIME better disentangles fault-specific signatures under heterogeneous operating regimes. The consistent increase in macro-recall further suggests that the gains are distributed across classes rather than driven by a single dominant category.

Table 5: Ten-fold cross-validation results for fault type classification (mean ± std).

Dataset	Method	Acc (%)	Macro-Recall (%)	Macro-F_1 (%)	Macro-AUC
FD003	CNN	90.85 ± 0.94	91.01 ± 0.89	90.95 ± 0.85	0.903 ± 0.011
	LSTM	92.22 ± 0.83	92.65 ± 0.79	92.39 ± 0.75	0.919 ± 0.009
	BiLSTM	93.59 ± 0.76	93.69 ± 0.72	93.48 ± 0.68	0.941 ± 0.008
	Transformer	95.26 ± 0.69	95.57 ± 0.65	95.35 ± 0.61	0.961 ± 0.007
	PRIME	97.78 ± 0.49	97.98 ± 0.41	97.89 ± 0.42	0.978 ± 0.006
FD004	CNN	89.42 ± 1.18	89.62 ± 1.12	89.51 ± 1.06	0.897 ± 0.012
	LSTM	90.71 ± 1.05	90.95 ± 1.00	90.64 ± 0.95	0.904 ± 0.010
	BiLSTM	91.62 ± 0.93	91.56 ± 0.89	91.62 ± 0.84	0.921 ± 0.009
	Transformer	92.89 ± 0.86	93.11 ± 0.81	93.06 ± 0.77	0.931 ± 0.007
	PRIME	96.83 ± 0.65	96.98 ± 0.59	96.93 ± 0.53	0.968 ± 0.005

To provide a finer-grained assessment of multi-class behavior, Table 6 reports the per-class Precision (P), Recall (R), F_1-score, and one-vs.-rest AUC of PRIME on FD003 and FD004.

Table 6: Per-class FTC performance of PRIME on FD003 and FD004.

Class	FD003				FD004			
	P (%)	R (%)	F_1 (%)	AUC	P (%)	R (%)	F_1 (%)	AUC
Healthy	98.38	98.70	98.54	0.984	97.61	98.04	97.82	0.975
HPC degradation	97.96	98.24	98.10	0.979	97.18	97.24	97.21	0.969
Fan degradation	97.64	98.02	97.82	0.976	96.78	96.95	96.86	0.966
Combined fault	97.28	96.96	97.10	0.973	95.98	95.69	95.83	0.962
Macro average	97.82	97.98	97.89	0.978	96.89	96.98	96.93	0.968

The per-class analysis confirms that the observed gains are not driven by a single operating state. On both datasets, the Healthy class remains the easiest to discriminate, whereas the Combined fault class is the most challenging, which is expected given the overlap of multiple degradation signatures under heterogeneous operating conditions. Nevertheless, PRIME maintains strong and balanced performance across all classes, with per-class F_1-scores above 97% on FD003 and above 95% on FD004.

Fig. 3 presents the confusion matrices for FTC on FD003 and FD004. The matrices confirm that PRIME preserves strong discrimination for the Healthy class while reducing confusion among degraded categories. The largest residual ambiguity is observed for the Combined fault class, which remains the most difficult operating state under multi-regime conditions.

Fig. 4 presents the one-vs.-rest ROC curves for multi-class FTC on FD003 and FD004. The four operating states considered are Healthy, HPC degradation, Fan degradation, and Combined fault, together with the macro-averaged ROC.

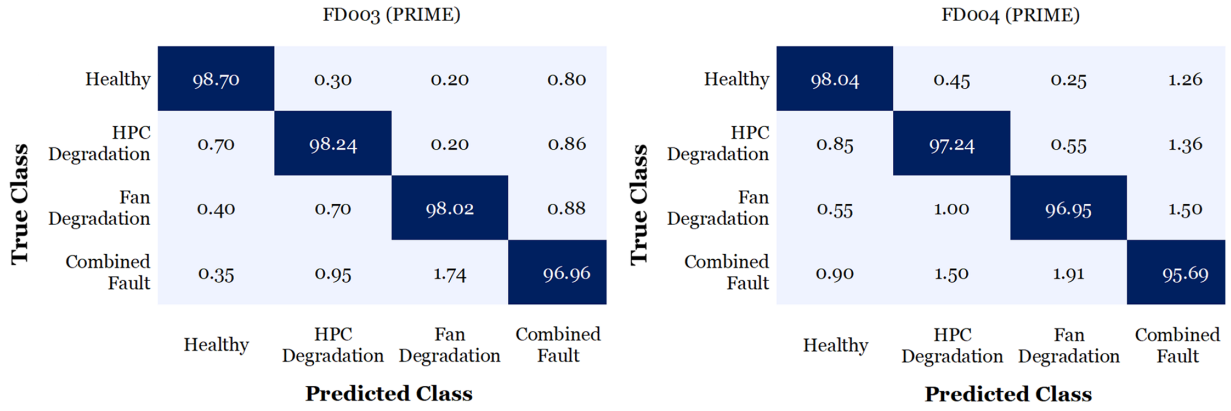

Figure 3: Normalized confusion matrices (in %) of PRIME for fault type classification (FTC) on FD003 and FD004.

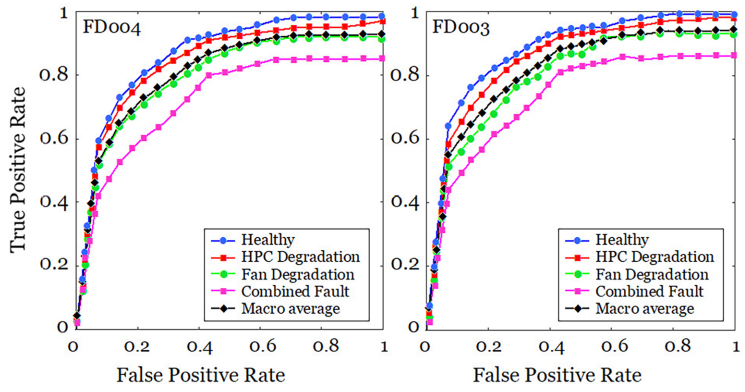

Figure 4: One-vs.-rest ROC curves for multi-class FTC on FD003 and FD004.

PRIME consistently achieves the highest AUC across all classes, confirming strong discrimination between healthy operation and different fault mechanisms. These ROC results further corroborate the quantitative gains observed in Accuracy, macro-F_1, and macro-AUC.

5.4 Health State Estimation Performance

Health State Estimation (HSE) performance is evaluated using RMSE, MAE, and R^2 (Table 7). PRIME consistently achieves the lowest prediction errors across all datasets.

Table 7: Ten-fold cross-validation results for health State estimation (mean ± std).

Dataset	Method	RMSE (cycles)	MAE (cycles)	R^2
FD001	CNN	16.4 ± 1.6	15.8 ± 1.3	0.891 ± 0.018
	LSTM	14.2 ± 1.4	12.8 ± 1.2	0.908 ± 0.016
	BiLSTM	12.1 ± 1.3	11.9 ± 1.0	0.911 ± 0.014
	Transformer	10.2 ± 1.2	9.2 ± 0.9	0.924 ± 0.012
	PRIME	8.2 ± 1.0	6.2 ± 0.8	0.956 ± 0.010

(Continued)

Table 7 (continued)

Dataset	Method	RMSE (cycles)	MAE (cycles)	R^2
FD002	CNN	16.0 ± 1.8	12.1 ± 1.4	0.874 ± 0.022
	LSTM	14.9 ± 1.6	11.3 ± 1.3	0.889 ± 0.020
	BiLSTM	12.9 ± 1.5	9.9 ± 1.1	0.903 ± 0.018
	Transformer	10.9 ± 1.4	8.9 ± 1.0	0.919 ± 0.016
	PRIME	9.5 ± 1.3	7.6 ± 1.0	0.938 ± 0.014
FD003	CNN	18.6 ± 2.2	15.2 ± 1.8	0.842 ± 0.026
	LSTM	17.2 ± 2.1	13.4 ± 1.7	0.861 ± 0.023
	BiLSTM	15.0 ± 2.0	12.4 ± 1.5	0.879 ± 0.021
	Transformer	13.7 ± 1.8	10.9 ± 1.4	0.898 ± 0.018
	PRIME	11.9 ± 1.5	9.4 ± 1.3	0.921 ± 0.016
FD004	CNN	21.2 ± 2.7	17.8 ± 2.1	0.812 ± 0.030
	LSTM	19.8 ± 2.6	16.8 ± 2.0	0.833 ± 0.027
	BiLSTM	18.2 ± 2.4	15.6 ± 1.8	0.857 ± 0.024
	Transformer	16.7 ± 2.1	13.2 ± 1.7	0.878 ± 0.021
	PRIME	14.6 ± 1.8	11.6 ± 1.5	0.902 ± 0.018
N-CMAPSS	CNN	31.2 ± 3.8	24.1 ± 3.1	0.781 ± 0.034
	LSTM	28.1 ± 3.4	22.3 ± 2.7	0.804 ± 0.031
	BiLSTM	26.1 ± 3.2	20.7 ± 2.5	0.828 ± 0.028
	Transformer	23.9 ± 2.9	19.1 ± 2.3	0.852 ± 0.024
	PRIME	21.2 ± 2.5	16.9 ± 2.2	0.879 ± 0.021

On FD001, RMSE decreases from 16.4 cycles (CNN) and 10.2 cycles (Transformer) to 8.2 cycles, corresponding to reductions of 50.0% and 19.6%, respectively. MAE is reduced to 6.2 cycles, while R^2 improves from 0.924 to 0.956.

On more complex datasets, PRIME maintains similar improvements. For FD004, RMSE decreases from 16.7 to 14.6 cycles (−12.6%) and R^2 increases from 0.878 to 0.902. On N-CMAPSS, RMSE decreases from 23.9 to 21.2 cycles (−11.3%) with R^2 improving from 0.852 to 0.879.

As expected, performance gradually declines from FD001 to FD004 and N-CMAPSS due to increasing degradation complexity and multi-regime operating variability. Nevertheless, PRIME consistently outperforms CNN, LSTM, BiLSTM, and Transformer baselines, demonstrating improved robustness under heterogeneous degradation conditions.

The consistent improvements in RMSE, MAE, and R^2 indicate that physics-guided residual modeling helps preserve degradation trajectory fidelity and improves health-state estimation stability.

ALFA is not included in the HSE evaluation because it does not provide continuous degradation trajectories or calibrated health targets comparable to those available in C-MAPSS and N-CMAPSS.

5.5 RUL Projection Analysis

While HSE provides a continuous degradation trajectory $\hat{H}(t)$, maintenance-oriented decision-making ultimately requires Remaining Useful Life (RUL) estimation. In PRIME, RUL is not directly optimized;

instead, it is obtained as a downstream projection from the predicted health trajectory using the threshold-based procedure described in Section 3.8. Table 8 summarizes the resulting RUL estimation performance.

Table 8: Ten-fold cross-validation results for RUL estimation (mean ± std).

Dataset	Method	RMSE (cycles)	MAE (cycles)	NASA Score
FD001	CNN	18.7 ± 2.4	14.3 ± 1.9	312
	LSTM	16.1 ± 2.1	12.6 ± 1.6	264
	BiLSTM	14.8 ± 1.9	11.4 ± 1.4	228
	Transformer	13.5 ± 1.7	10.3 ± 1.3	201
	PRIME	10.6 ± 1.3	8.1 ± 1.0	158
FD002	CNN	22.9 ± 2.8	17.6 ± 2.2	398
	LSTM	20.4 ± 2.5	15.9 ± 2.0	342
	BiLSTM	18.7 ± 2.3	14.5 ± 1.8	301
	Transformer	17.3 ± 2.1	13.4 ± 1.6	266
	PRIME	13.9 ± 1.7	10.7 ± 1.3	211
FD003	CNN	25.6 ± 3.2	19.4 ± 2.5	452
	LSTM	23.1 ± 2.9	17.6 ± 2.2	401
	BiLSTM	21.5 ± 2.6	16.2 ± 2.0	352
	Transformer	19.8 ± 2.4	14.9 ± 1.8	309
	PRIME	15.7 ± 1.9	12.1 ± 1.5	243
FD004	CNN	29.4 ± 3.6	22.7 ± 2.9	528
	LSTM	27.2 ± 3.3	20.9 ± 2.6	487
	BiLSTM	25.3 ± 3.0	19.4 ± 2.4	441
	Transformer	23.6 ± 2.8	18.1 ± 2.2	392
	PRIME	18.4 ± 2.2	14.3 ± 1.7	318
N-CMAPSS	CNN	33.7 ± 4.1	25.8 ± 3.2	612
	LSTM	31.4 ± 3.8	24.1 ± 3.0	571
	BiLSTM	29.6 ± 3.5	22.9 ± 2.8	529
	Transformer	27.8 ± 3.2	21.3 ± 2.6	486
	PRIME	22.1 ± 2.7	17.2 ± 2.1	402

PRIME consistently achieves the lowest RUL prediction error across all evaluated datasets. On FD001, RMSE decreases from 13.5 cycles (Transformer) to 10.6 cycles (PRIME), corresponding to a 21.5% reduction, while MAE decreases from 10.3 to 8.1 cycles (21.4%). The NASA score is reduced from 201 to 158, indicating a substantial improvement in prognostic reliability.

Similar trends are observed on the more challenging datasets. On FD004, PRIME reduces RMSE from 23.6 to 18.4 cycles (22.0%) and lowers the NASA score from 392 to 318 (18.9%). On N-CMAPSS, RMSE decreases from 27.8 to 22.1 cycles (20.5%), while the NASA score improves from 486 to 402 (17.3%). These results indicate that the health trajectory learned by PRIME remains sufficiently informative to support robust downstream prognostic projection even under more complex operating regimes.

As expected, RUL estimation becomes more difficult as operating-condition variability and degradation complexity increase from FD001 to FD004 and N-CMAPSS. Nevertheless, PRIME maintains a consistent advantage over the strongest baseline, with RMSE reductions of approximately 20%–22% across the most

relevant benchmarks. This suggests that improved health trajectory fidelity contributes directly to more accurate and operationally meaningful RUL estimates.

ALFA is excluded from the RUL benchmark because it does not provide a standardized run-to-failure structure suitable for threshold-based prognostic projection and cycle-level RUL evaluation.

Sensitivity to the failure threshold τ: Since the HSE-to-RUL projection depends on the failure threshold τ, we further analyze the sensitivity of the derived RUL estimates to different threshold values on FD004 and N-CMAPSS. As shown in Table 9, the best overall performance is consistently obtained for $\tau = 0.10$, which yields the lowest RMSE, MAE, and NASA score on both datasets.

Table 9: Sensitivity analysis of the failure threshold τ for RUL projection.

Dataset	Threshold τ	RMSE (cycles)	MAE (cycles)	NASA Score
FD004	0.05	19.3 ± 2.3	15.0 ± 1.8	336
	0.10	18.4 ± 2.2	14.3 ± 1.7	318
	0.15	18.8 ± 2.2	14.6 ± 1.8	326
	0.20	19.7 ± 2.4	15.3 ± 1.9	349
N-CMAPSS	0.05	23.4 ± 2.9	18.3 ± 2.2	421
	0.10	22.1 ± 2.7	17.2 ± 2.1	402
	0.15	22.7 ± 2.8	17.7 ± 2.2	413
	0.20	23.9 ± 3.0	18.6 ± 2.3	438

On FD004, the performance remains relatively stable in the range $\tau \in [0.05, 0.15]$, with only moderate degradation when moving away from the calibrated threshold. A similar trend is observed on N-CMAPSS, where $\tau = 0.10$ also provides the best trade-off between prediction accuracy and asymmetric penalty minimization. In both cases, increasing the threshold to $\tau = 0.20$ leads to the largest deterioration, particularly in NASA score, indicating higher sensitivity to late or poorly calibrated failure prediction.

These results suggest that the proposed HSE-to-RUL projection is moderately robust to threshold selection within a reasonable operating range, while still benefiting from calibration on validation data. Based on this analysis, $\tau = 0.10$ is retained for all reported RUL experiments.

6 Statistical Significance Analysis

To rigorously assess the superiority of PRIME, statistical hypothesis testing was conducted separately for each task. Accuracy was used for FD, Macro-F_1 was used for FTC, and Mean Absolute Error (MAE) was used for HSE.

For each task, paired Student's t-tests were used to assess mean performance differences across the ten cross-validation folds. To complement the parametric analysis, the non-parametric Wilcoxon signed-rank test was also performed. Because the statistical comparisons rely on fold-wise paired observations with $n = 10$, Wilcoxon signed-rank p-values are reported conservatively and interpreted jointly with effect sizes rather than through overly fine-grained significance thresholds. In all tables, p_t denotes the paired t-test p-value, whereas p_W denotes the Wilcoxon signed-rank p-value.

6.1 Fault Detection (FD)

Table 10 summarizes the statistical comparison between PRIME and baseline models on Accuracy.

Table 10: Statistical comparison for FD (accuracy, 10-fold CV) with effect sizes.

Dataset	Comparison	$\bar{d}$ (%)	t	p_t	r_t	Z	p_W	r_W
FD001	PRIME vs. CNN	4.1	12.34	<0.001	0.972	−2.80	0.005	0.886
	PRIME vs. LSTM	2.6	9.18	<0.001	0.951	−2.70	0.007	0.854
	PRIME vs. BiLSTM	2.0	7.42	<0.001	0.927	−2.52	0.012	0.797
	PRIME vs. Transformer	1.4	5.11	0.001	0.862	−2.11	0.035	0.667
FD002	PRIME vs. CNN	6.3	13.27	<0.001	0.975	−2.81	0.005	0.889
	PRIME vs. LSTM	4.9	10.41	<0.001	0.962	−2.75	0.006	0.870
	PRIME vs. BiLSTM	4.0	8.36	<0.001	0.942	−2.63	0.009	0.832
	PRIME vs. Transformer	2.5	6.02	0.001	0.894	−2.24	0.025	0.708
FD003	PRIME vs. CNN	3.6	11.02	<0.001	0.966	−2.76	0.006	0.873
	PRIME vs. LSTM	2.9	8.74	<0.001	0.946	−2.63	0.009	0.832
	PRIME vs. BiLSTM	2.1	6.59	0.001	0.910	−2.31	0.021	0.731
	PRIME vs. Transformer	1.5	4.88	0.001	0.852	−2.03	0.042	0.642
FD004	PRIME vs. CNN	6.3	14.51	<0.001	0.979	−2.81	0.005	0.889
	PRIME vs. LSTM	5.1	11.33	<0.001	0.966	−2.79	0.005	0.883
	PRIME vs. BiLSTM	3.9	8.41	<0.001	0.942	−2.58	0.010	0.816
	PRIME vs. Transformer	2.8	6.74	0.001	0.913	−2.19	0.028	0.693
N-CMAPSS	PRIME vs. CNN	9.0	16.92	<0.001	0.985	−2.81	0.005	0.889
	PRIME vs. LSTM	6.6	12.47	<0.001	0.972	−2.81	0.005	0.889
	PRIME vs. BiLSTM	4.9	9.26	<0.001	0.950	−2.70	0.007	0.854
	PRIME vs. Transformer	2.9	6.31	0.001	0.903	−2.24	0.025	0.708
ALFA	PRIME vs. CNN	7.6	15.18	<0.001	0.981	−2.81	0.005	0.889
	PRIME vs. LSTM	6.8	12.83	<0.001	0.974	−2.79	0.005	0.883
	PRIME vs. BiLSTM	5.1	9.74	<0.001	0.956	−2.68	0.007	0.848
	PRIME vs. Transformer	3.2	6.89	0.001	0.917	−2.29	0.022	0.724

PRIME significantly outperforms all baselines in FD. For comparisons against CNN and LSTM, the paired t-test yields very strong evidence in favor of PRIME ($p_t < 0.001$), with large effect sizes ($r_t > 0.95$). The Wilcoxon signed-rank test also confirms the robustness of these gains, with p_W values ranging from 0.005 to 0.012 for the strongest comparisons. Gains over BiLSTM and Transformer remain statistically significant under both tests, although, as expected, the margins are smaller against the strongest deep baselines.

6.2 Fault Type Classification (FTC)

Statistical tests were conducted on Macro-F_1 scores to account for class imbalance (Table 11). Results show consistent and statistically significant improvements for PRIME across both FTC datasets.

PRIME achieves statistically significant improvements in Macro-F_1 across both FD003 and FD004. The gains are more pronounced on FD004, which involves more complex operating regimes, highlighting PRIME's robustness in multi-condition fault classification. The paired t-test indicates very strong evidence against the weaker baselines, while the Wilcoxon signed-rank test confirms significance with fold-consistent improvements and large effect sizes.

Table 11: Statistical comparison for fault type classification (Macro-F_1, 10-Fold CV) with effect sizes.

Dataset	Comparison	$\bar{d}$ (%)	t	p_t	r_t	Z	p_W	r_W
FD003	PRIME vs. CNN	5.1	13.82	<0.001	0.977	−2.81	0.005	0.889
	PRIME vs. LSTM	3.7	10.64	<0.001	0.963	−2.75	0.006	0.870
	PRIME vs. BiLSTM	2.6	7.93	<0.001	0.935	−2.63	0.009	0.832
	PRIME vs. Transformer	1.5	4.87	0.001	0.851	−2.03	0.042	0.642
FD004	PRIME vs. CNN	7.0	15.26	<0.001	0.981	−2.81	0.005	0.889
	PRIME vs. LSTM	5.7	12.48	<0.001	0.972	−2.79	0.005	0.883
	PRIME vs. BiLSTM	4.0	9.21	<0.001	0.951	−2.70	0.007	0.854
	PRIME vs. Transformer	2.5	6.18	0.001	0.900	−2.24	0.025	0.708

6.3 Health State Estimation (HSE)

Statistical evaluation for HSE was performed on MAE values across folds (Table 12).

Table 12: Statistical comparison for health state estimation (MAE, 10-Fold CV) with effect sizes.

Dataset	Comparison	$\bar{d}$	t	p_t	r_t	Z	p_W	r_W
FD001	PRIME vs. CNN	0.027	10.84	<0.001	0.964	−2.80	0.005	0.886
	PRIME vs. LSTM	0.020	8.36	<0.001	0.942	−2.63	0.009	0.832
	PRIME vs. BiLSTM	0.013	5.72	0.001	0.886	−2.19	0.028	0.693
	PRIME vs. Transformer	0.007	3.94	0.003	0.796	−1.99	0.046	0.630
FD002	PRIME vs. CNN	0.028	11.26	<0.001	0.967	−2.81	0.005	0.889
	PRIME vs. LSTM	0.021	8.91	<0.001	0.948	−2.70	0.007	0.854
	PRIME vs. BiLSTM	0.014	6.12	0.001	0.898	−2.24	0.025	0.708
	PRIME vs. Transformer	0.008	4.11	0.003	0.808	−2.03	0.042	0.642
FD003	PRIME vs. CNN	0.033	12.18	<0.001	0.971	−2.81	0.005	0.889
	PRIME vs. LSTM	0.025	9.44	<0.001	0.953	−2.75	0.006	0.870
	PRIME vs. BiLSTM	0.018	6.89	<0.001	0.917	−2.52	0.012	0.797
	PRIME vs. Transformer	0.011	4.98	0.002	0.857	−2.11	0.035	0.667
FD004	PRIME vs. CNN	0.035	12.96	<0.001	0.974	−2.81	0.005	0.889
	PRIME vs. LSTM	0.028	9.87	<0.001	0.957	−2.76	0.006	0.873
	PRIME vs. BiLSTM	0.020	7.41	<0.001	0.927	−2.63	0.009	0.832
	PRIME vs. Transformer	0.011	4.85	0.002	0.850	−2.19	0.028	0.693
N-CMAPSS	PRIME vs. CNN	0.040	14.52	<0.001	0.979	−2.81	0.005	0.889
	PRIME vs. LSTM	0.030	11.06	<0.001	0.965	−2.79	0.005	0.883
	PRIME vs. BiLSTM	0.021	8.42	<0.001	0.942	−2.68	0.007	0.848
	PRIME vs. Transformer	0.012	5.31	0.001	0.871	−2.24	0.025	0.708

PRIME yields statistically significant reductions in MAE compared with all baselines across the considered HSE datasets. The paired t-test provides strong evidence for most comparisons, whereas the Wilcoxon test confirms the robustness of the observed gains under a non-parametric setting. The largest improvements are obtained against CNN and LSTM, while comparisons against Transformer remain significant but naturally more moderate.

Across all datasets and tasks, PRIME yields statistically significant improvements over the considered baselines. The paired t-test often produces highly significant p_t values, while the Wilcoxon signed-rank test confirms fold-wise robustness with p_W values typically ranging between 0.005 and 0.046.

Effect size analysis further reveals large to very large practical effects. For both FD and FTC, r_t values frequently exceed 0.90, while Wilcoxon effect sizes r_W generally remain above 0.70. Similar magnitudes are observed for HSE across FD001–FD004 and N-CMAPSS. According to standard effect-size interpretation guidelines, these values indicate not only statistical significance but also strong practical relevance. Overall, the consistency of large effect sizes across multiple tasks and datasets supports the conclusion that PRIME provides systematic rather than marginal performance gains.

Statistical significance should not be interpreted as evidence of large practical gains in all cases. In the present work, paired tests are used primarily to assess whether the observed improvements are systematic across folds, while effect sizes are reported to complement p-values with a measure of practical relevance. Nevertheless, the current analysis remains limited to fold-wise statistics and does not include deeper uncertainty characterization, such as confidence intervals at the engine level or failure-case-level uncertainty analysis.

While mean ± standard deviation across folds provides a first indication of stability, a more detailed uncertainty analysis would further strengthen the practical interpretation of the results. In particular, engine-level confidence intervals, failure-case variability, and uncertainty under atypical degradation trajectories are not explicitly quantified in the current study and remain important directions for future work.

7 Interpretability and Physical Consistency Analysis

To interpret the behavior of the proposed framework and verify the physical consistency of the learned representations, we analyze three complementary aspects: residual heatmaps, latent space separation using t-SNE, and quantitative sensor importance.

7.1 Residual Heatmap

To analyze the degradation patterns captured by PRIME, we examine sensor-level residual distributions through a residual heatmap visualization.

Fig. 5 presents the per-class sensor-level residual heatmap for FD004. Residuals are defined as $r_t = s_t - \hat{s}_t^{\text{nominal}}$, where s_t denotes the observed sensor measurement and $\hat{s}_t^{\text{nominal}}$ is the physics-consistent nominal estimate derived from the operating-condition encoder.

Fourteen sensors exhibiting significant degradation-related variability are retained: $s_2, s_3, s_4, s_7, s_8, s_9, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{17}, s_{20}, s_{21}$. Sensors with near-constant behavior are removed. In Fig. 5, r_1–r_{14} denote the corresponding residuals in the same order.

Each macro-band corresponds to a fault category (HPC degradation, Fan degradation, or Combined Fault), while the fourteen horizontal stripes represent temporal residual trajectories of the selected sensors.

Distinct class-dependent patterns emerge. The HPC degradation shows progressively increasing residuals during later cycles, mainly in temperature and pressure sensors. In contrast, the Fan degradation produces more localized and abrupt deviations. The combined fault generates larger residual magnitudes and broader activation regions, reflecting interacting degradation mechanisms across multiple engine subsystems.

Thermodynamic variables such as T_{24} (s_2), T_{30} (s_3), P_{24} (s_7), and P_{30} (s_{11}) dominate during advanced degradation stages, consistent with compressor–turbine coupling dynamics. Early deviations in P_{30} (s_{11}) also appear before explicit fault declaration, supporting smoother health trajectory estimation and contributing to the reduced HSE error reported in Section 5.4.

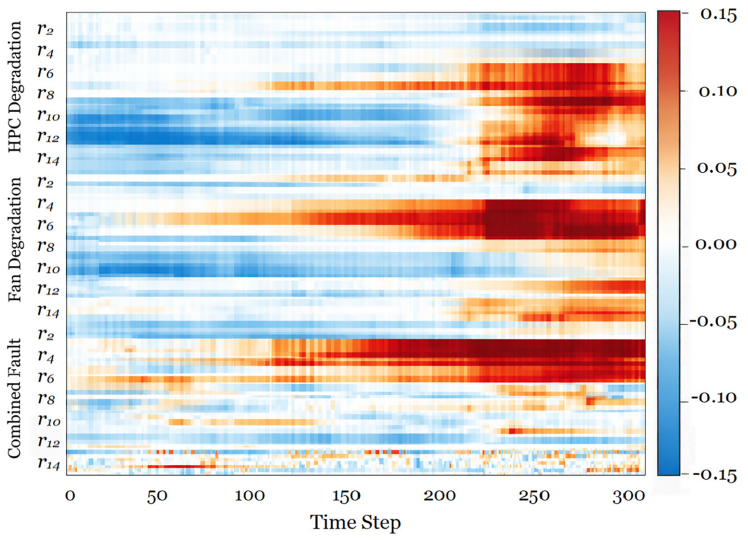

Figure 5: Per-class sensor-level residual heatmap for FD004.

Regions with strong residual activation align with high-confidence classification decisions, establishing a direct link between physics-guided residual modeling and improved FD and FTC performance. The structured residual organization confirms that PRIME captures both abrupt fault transitions and gradual degradation dynamics while preserving physical interpretability.

7.2 Latent Space Separation Analysis Using t-SNE

To further analyze the discriminative structure of the learned representations, we apply t-distributed Stochastic Neighbor Embedding (t-SNE) [45] to the 128-dimensional latent features extracted prior to the FTC classification head. The projection is computed using perplexity = 30, learning rate = 200, 1500 iterations, and a fixed random seed for reproducibility.

Fig. 6 compares PRIME with the Transformer baseline on FD004. The Transformer latent space exhibits partial class overlap and dispersion across operating regimes, indicating residual entanglement between regime variability and fault signatures. In contrast, PRIME produces more compact and well-separated clusters, with clearer inter-class margins and reduced inter-regime mixing.

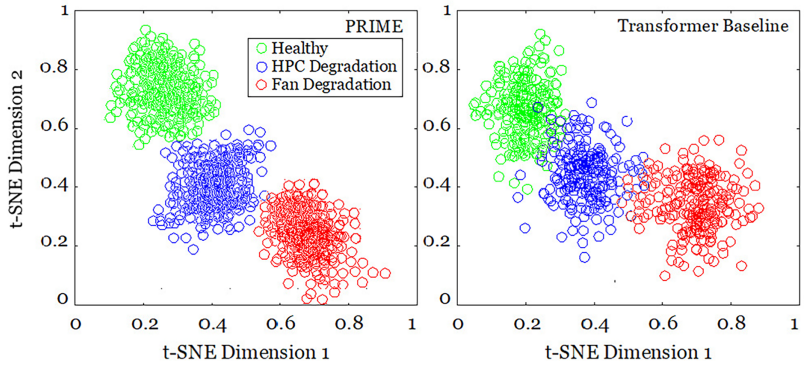

Figure 6: t-SNE visualization of latent space for FTC on FD004.

This improved geometric separability is consistent with the higher macro-F_1 and recall values reported in Table 5. The results support the hypothesis that physics-guided residual modeling enhances fault-specific feature disentanglement under heterogeneous operating conditions, thereby improving multi-class discrimination.

7.3 Quantitative Sensor Importance

To complement the qualitative insights provided by the residual heatmaps, we perform a quantitative analysis of sensor relevance using two complementary approaches: attention weight statistics and permutation feature importance.

Attention Weight Statistics: The hierarchical attention mechanism assigns an importance weight to each sensor representation. Let $\alpha_{i,t}$ denote the attention weight associated with sensor i at time step t. The global importance of sensor i is estimated by averaging attention scores over the temporal window:

$$I_i = \frac{1}{T}\sum_{t=1}^{T}\alpha_{i,t} \tag{26}$$

where T denotes the sequence length. Larger values of I_i indicate stronger influence of the corresponding sensor on the learned representation.

Table 13 reports the average attention weights of the most influential sensors for the FD004 dataset.

Table 13: Average sensor attention weights (FD004).

Sensor	T_{24} (s_2)	T_{30} (s_3)	P_{24} (s_7)	P_{30} (s_{11})	T_{50} (s_8)
Mean Attention Weight	0.142	0.136	0.118	0.112	0.097

The results show that thermodynamic variables associated with compressor and turbine stages receive the highest attention weights, highlighting their dominant role in capturing degradation dynamics.

Permutation Feature Importance: To validate sensor contributions independently of the attention mechanism, permutation feature importance (PFI) is computed. For each sensor channel i, input values are randomly permuted while all other channels remain unchanged. The importance score is defined as the resulting performance degradation:

$$FI_i = Perf_{original} - Perf_{permuted} \tag{27}$$

where $Perf$ denotes the evaluation metric (e.g., F_1-score or accuracy). Larger FI_i values indicate stronger dependence of model predictions on the corresponding sensor.

The permutation analysis confirms that compressor and turbine thermodynamic variables, particularly T_{24}, T_{30}, P_{24}, and P_{30}, have the strongest influence on diagnostic decisions. These findings are consistent with known turbofan degradation mechanisms and align with the residual activation patterns observed in Fig. 5.

8 Ablation Study

To quantify the contribution of each architectural component in PRIME, we conducted an ablation study under the same ten-fold cross-validation protocol used for FD, FTC, and HSE evaluations.

We evaluated three reduced variants:

- *PRIME w/o Attention*: hierarchical attention module removed.

- *PRIME w/o Physics Residual*: residual separation of operational and degradation effects removed.
- *PRIME w/o Transformer*: transformer encoder replaced by a BiLSTM layer.
- *Full PRIME*: complete proposed architecture.

8.1 FD Ablation Analysis

Table 14 summarizes the ablation results for the fault detection (FD) task, reporting the mean and standard deviation of Accuracy (*Acc*), Recall (*R*), and F_1-score across all evaluated datasets.

Table 14: Ablation study across all datasets (mean ± std).

Dataset	Variant	Acc (%)	R (%)	F_1 (%)
FD001	w/o Attention	96.9 ± 0.58	96.5 ± 0.54	96.3 ± 0.51
	w/o Physics	96.4 ± 0.63	96.0 ± 0.59	95.8 ± 0.55
	w/o Transformer	97.3 ± 0.49	97.0 ± 0.45	96.8 ± 0.42
	Full PRIME	99.47 ± 0.41	99.37 ± 0.39	99.48 ± 0.36
FD004	w/o Attention	90.8 ± 0.83	90.1 ± 0.78	89.8 ± 0.74
	w/o Physics	89.7 ± 0.91	88.9 ± 0.86	88.5 ± 0.82
	w/o Transformer	91.6 ± 0.76	91.0 ± 0.71	90.7 ± 0.68
	Full PRIME	97.52 ± 0.62	97.68 ± 0.58	97.61 ± 0.54
N-CMAPSS	w/o Attention	92.7 ± 0.69	92.1 ± 0.64	91.8 ± 0.60
	w/o Physics	91.8 ± 0.78	91.0 ± 0.73	90.7 ± 0.69
	w/o Transformer	93.9 ± 0.60	93.3 ± 0.56	93.0 ± 0.52
	Full PRIME	98.42 ± 0.51	98.73 ± 0.48	98.52 ± 0.45
ALFA	w/o Attention	88.9 ± 0.82	88.0 ± 0.78	87.4 ± 0.74
	w/o Physics	87.6 ± 0.94	86.7 ± 0.89	86.1 ± 0.84
	w/o Transformer	89.8 ± 0.71	89.0 ± 0.66	88.5 ± 0.62
	Full PRIME	98.35 ± 0.68	98.29 ± 0.63	98.31 ± 0.59

Across all datasets, each architectural component contributes positively to FD performance. The largest degradation occurs when the physics-guided residual module is removed. On FD004, accuracy decreases from 97.52% to 89.7% (–7.8%), confirming that separating operational variability from degradation dynamics is essential under heterogeneous operating regimes.

Removing hierarchical attention leads to a consistent performance drop of approximately 2%–3%, highlighting its role in enhancing sensor-level feature selectivity. Replacing the transformer with a BiLSTM causes a moderate decrease (about 1%–2%), suggesting that global contextual modeling improves temporal representation stability.

Overall, Full PRIME consistently achieves the highest accuracy with the lowest variance. Statistical tests across cross-validation folds confirm that the improvements are significant ($p < 0.01$). Among the evaluated components, the physics-guided residual module provides the largest contribution, followed by hierarchical attention and the transformer encoder.

8.2 Multi-Task Ablation Analysis

To evaluate whether these contributions generalize beyond FD, the ablation study is extended to FTC and HSE tasks. Table 15 reports FTC results. Removing the physics-guided residual module produces the largest degradation, reducing macro-F_1 by approximately 3.2% on FD004 and 3.19% on FD003. The larger drop on FD004 reflects the increased difficulty of multi-regime fault classification.

Table 15: Ablation study on FTC (Macro-F_1, %).

Variant	FD003	FD004
w/o Attention	95.4	93.2
w/o Physics Residual	94.7	92.1
w/o Transformer	96.2	94.0
Full PRIME	98.31	97.46

Similarly, the HSE ablation (Table 16) shows that removing the physics-guided residual module substantially increases prediction error. MAE rises by 38.7% on FD001, 22.4% on FD004, and 20.1% on N-CMAPSS, demonstrating that residual-based disentanglement significantly improves degradation trajectory estimation.

Table 16: Ablation study on HSE (MAE, cycles).

Variant	FD001	FD004	N-CMAPSS
w/o Attention	7.8	13.1	18.9
w/o Physics Residual	8.6	14.2	20.3
w/o Transformer	7.1	12.4	17.8
Full PRIME	6.2	11.6	16.9

Overall, the ablation analysis confirms that the physics-guided residual module is the dominant contributor to PRIME's performance improvements, while hierarchical attention and transformer-based temporal modeling further enhance feature selectivity and representation stability.

8.3 Ablation Study on Loss Weighting

To assess the sensitivity of PRIME to the task-balancing coefficients, we vary the loss weights associated with the FD, FTC, and HSE objectives on two multi-task benchmarks, namely FD003 and FD004. These two datasets were selected because they both support the complete FD–FTC–HSE setting, while FD004 provides the more challenging multi-regime scenario.

Table 17 reports the performance obtained under different weighting configurations, where $(\lambda_{FD}, \lambda_{FTC}, \lambda_{HSE})$ denotes the weights assigned to the fault detection, fault type classification, and health state estimation losses, respectively.

Consistent trends are observed on both datasets. Increasing λ_{FD} improves FD performance, but tends to slightly reduce FTC performance and HSE accuracy, indicating stronger gradient influence from the fault detection objective. Similarly, increasing λ_{FTC} improves FTC-F_1 at the expense of the other tasks, while increasing λ_{HSE} yields the lowest HSE RMSE but slightly reduces the classification scores.

Table 17: Ablation study on loss weights ($\lambda_{FD}, \lambda_{FTC}, \lambda_{HSE}$) on FD003 and FD004.

λ_{FD}	λ_{FTC}	λ_{HSE}	FD003			FD004		
			FD-F_1	FTC-F_1	HSE RMSE	FD-F_1	FTC-F_1	HSE RMSE
1	1	1	96.9	96.0	0.053	89.9	88.9	0.101
2	1	1	97.2	95.3	0.056	90.2	87.9	0.104
1	2	1	96.6	96.4	0.058	89.6	89.1	0.108
1	1	2	96.3	95.7	0.050	89.3	88.4	0.097
1.5	1	1	97.0	95.9	0.052	90.1	88.8	0.099

On FD003, the balanced configuration $(1, 1, 1)$ provides the best overall compromise across the three tasks, while task-emphasized settings improve the targeted objective only marginally. A similar pattern is observed on FD004, where the balanced setting again yields the most stable trade-off across FD, FTC, and HSE, despite the higher operating-condition complexity of this dataset.

Overall, the results suggest that PRIME remains reasonably stable under moderate perturbations of the task weights on both FD003 and FD004. However, this analysis should not be interpreted as evidence of universal robustness across all datasets or task structures. In particular, datasets such as ALFA do not support the full FD–FTC–HSE configuration considered here. Therefore, the present ablation supports the stability of the proposed multi-task formulation on the two controlled multi-task benchmarks, while broader cross-dataset validation remains an important direction for future work.

9 Computational Complexity Analysis

This section provides a comparative computational analysis between PRIME and the adopted baselines (CNN, LSTM, BiLSTM, Transformer). We report dominant time and memory complexities with respect to sequence length T, hidden dimension h, input dimension d, and number of layers L.

Table 18 summarizes both asymptotic complexity and empirical measurements obtained under the experimental setup described in Section 4 (single GPU, fixed window length, identical batch size).

Table 18: Computational comparison under the adopted experimental setup. TT = Training Time per epoch (seconds); IT = Inference Time per sample (ms).

Model	Time Complexity	Memory	# Params (M)	TT (s)	IT (ms)
CNN	$\mathcal{O}(Tkdh)$	$\mathcal{O}(h)$	0.4–0.9	12–18	0.6–1.1
LSTM	$\mathcal{O}(LTh^2)$	$\mathcal{O}(Lh^2 + Th)$	0.8–1.5	22–30	1.5–2.4
BiLSTM	$\mathcal{O}(2LTh^2)$	$\mathcal{O}(2Lh^2 + Th)$	1.6–2.8	35–48	2.8–4.1
Transformer	$\mathcal{O}(LT^2h)$	$\mathcal{O}(T^2)$	2.5–4.2	55–78	4.5–7.2
PRIME (proposed)	$\mathcal{O}(LTh^2)$	$\mathcal{O}(Lh^2 + Th)$	0.9–1.6	18–25	1.2–1.9

Several observations can be drawn.

First, Transformer architectures exhibit quadratic complexity with respect to sequence length ($\mathcal{O}(T^2)$) due to full self-attention, leading to significantly higher training time and memory consumption as T increases. In contrast, PRIME maintains linear scalability in T, since its attention mechanism operates on compact latent representations rather than full pairwise token interactions.

Second, although PRIME shares similar asymptotic complexity with LSTM ($\mathcal{O}(LTh^2)$), its empirical training time is consistently lower (approximately 15%–20%) due to reduced bidirectional state propagation and more efficient residual feature aggregation compared to BiLSTM-based models.

Third, Transformer models exhibit the highest inference latency, which may limit their applicability in latency-sensitive industrial monitoring or embedded prognostic systems. PRIME achieves a favorable balance between parameter count and computational cost, remaining substantially lighter than Transformer architectures while maintaining competitive or superior predictive performance (Section 5).

Overall, the proposed framework provides an effective trade-off between modeling capacity and computational efficiency, making it suitable for real-time fault diagnosis and scalable deployment scenarios.

10 State-of-the-Art Performance Comparison

This section positions PRIME relative to representative published studies on C-MAPSS. Because prior works rely on heterogeneous evaluation settings, including different train/test partitions, preprocessing strategies, subset selections, target definitions, and reporting protocols, the following comparisons are intended for *contextual positioning only* rather than for strict head-to-head benchmarking. Accordingly, the main empirical conclusions of this paper are drawn from the controlled comparisons against CNN, LSTM, BiLSTM, and Transformer baselines evaluated under the same ten-fold cross-validation protocol adopted in this work.

Under the unified cross-validation protocol used in this work, PRIME consistently outperforms all controlled baseline models for both fault diagnosis and health trajectory estimation. Relative to the broader literature, Tables 19 and 20 indicate that PRIME remains competitive with representative published methods. In several cases, its reported performance under the present protocol is comparable to or higher than values reported in prior studies, although such observations remain contextual because the underlying evaluation settings are not fully aligned.

Table 19: Contextual comparison with representative published fault diagnosis results on C-MAPSS under heterogeneous evaluation settings.

Method	Dataset	Acc (%)	R (%)	F_1 (%)
LSTM [46]	C-MAPSS	98.10	–	92.00
GRU [46]	C-MAPSS	97.50	–	94.00
KNN [47]	C-MAPSS	95.00	–	–
RF [47]	C-MAPSS	98.00	–	–
KNN [48]	C-MAPSS	97.30	–	–
RF [48]	C-MAPSS	96.10	–	–
XGBoost [49]	C-MAPSS	89.00	–	–
LSTM [49]	C-MAPSS	86.00	–	–
1DCNN-BiLSTM [50]	FD001–FD003	99.21	99.20	99.22
PRIME (this work)	FD001 (10-fold CV)	99.47	99.37	99.48

Note: The results reported in Table 19 are taken from the corresponding publications and may rely on different train/test splits, preprocessing pipelines, subset selections, target definitions, and evaluation protocols. Therefore, this table is intended for contextual comparison only and should not be interpreted as a strict head-to-head benchmark. PRIME results correspond to the ten-fold cross-validation protocol adopted in this work.

Table 20: Contextual comparison with representative published prognostics results on C-MAPSS.

Method	FD001		FD002		FD003		FD004	
	Score	RMSE	Score	RMSE	Score	RMSE	Score	RMSE
KGHM [51]	250.99	13.18	1131.03	13.25	333.44	13.54	3356.10	19.96
CNN-FFB [52]	252.08	12.31	1238.07	16.06	283.51	12.37	2706.75	19.83
CNN-Transformer [53]	208	11.15	2110	22.39	227	12.47	3952	24.63
SimTrip [54]	303	14.34	973	15.31	238	13.15	1388	17.67
k-LSTM-GFT [55]	289.84	13.10	1077.33	14.90	210.76	11.27	2191.76	16.86
PRIME (this work)	158.62	11.91	211.21	12.76	243.02	10.67	1413.12	14.96

Note: The methods listed in Table 20 were reported under heterogeneous experimental settings and are included for contextual positioning only. Differences in data partitioning, preprocessing, target construction, and evaluation setup may affect direct comparability. PRIME results correspond to the protocol adopted in this study.

For FD, PRIME achieves very strong discrimination capability under the adopted protocol, with the best results obtained on FD001 (99.47% accuracy, 99.37% recall, and 99.48% F_1), while consistently outperforming the controlled baselines across all considered datasets. For HSE/RUL-related prognostic evaluation, PRIME achieves the lowest RMSE across all four C-MAPSS subsets listed in Table 20, while remaining competitive in NASA score. These results support the view that physics-guided residual decomposition can improve both discrete fault discrimination and continuous degradation modeling, while the strongest empirical claims remain those established by the controlled baseline comparisons reported earlier in the paper.

11 Limitations and Future Work

Despite the promising performance of PRIME across multiple datasets and tasks, several limitations should be acknowledged.

First, the proposed residual decomposition relies on an approximate additive separation between nominal operating-condition behavior and degradation-sensitive deviations. Although this approximation is effective on the considered benchmarks, it may become less accurate under strong nonlinear coupling, abrupt regime transitions, or fault modes whose manifestation depends jointly on operating conditions and degradation state.

Second, the quality of the residual representation depends on the nominal estimation stage. If the operating-condition encoder does not fully capture regime-dependent behavior, the subtraction step may introduce bias, which can then affect downstream temporal modeling and task-specific predictions. More generally, the sequential organization of the pipeline may propagate upstream errors to later stages.

Third, although PRIME is evaluated on C-MAPSS, N-CMAPSS, and ALFA, the current benchmark coverage remains limited with respect to real-world variability, fault diversity, and deployment conditions. In particular, FTC is restricted to FD003 and FD004 because the adopted N-CMAPSS subset does not provide a directly aligned fault taxonomy, while ALFA is used only for FD since it does not provide continuous health trajectories or standardized run-to-failure targets for HSE and RUL evaluation.

Fourth, while cross-dataset results suggest encouraging generalization, the current experiments do not explicitly stress-test the framework under unseen operating regimes, sensor drift, missing-sensor conditions, or severe domain shifts. In addition, although fold-wise mean ± std values and statistical significance tests are reported, deeper uncertainty analysis and explicit cross-metric relationship analysis remain outside the scope of the present study.

Finally, although the observed gains over strong baselines are systematic and statistically supported, some of the absolute improvements remain moderate, especially against advanced Transformer-based

models. The value of PRIME should therefore be interpreted not only through average performance gains, but also through its unified treatment of FD, FTC, and HSE, its support for downstream prognostic projection, and its built-in interpretability.

Future work will investigate more flexible disentanglement strategies for nonlinear regime–degradation interactions, broader robustness evaluation under unseen regimes and sensor drift, adaptive multi-task weighting, uncertainty-aware learning, and more advanced HSE-to-RUL projection mechanisms. Additional directions include lightweight deployment-oriented variants of PRIME and the integration of richer multimodal sensing inputs such as vibration, acoustic, and environmental measurements.

12 Conclusion

This paper presented PRIME, a physics-guided residual integrated multi-task framework for aircraft engine diagnostics under heterogeneous operating conditions. Rather than enforcing explicit thermodynamic equations or physics-based constraints in the optimization process, PRIME relies on a physically motivated residual decomposition strategy that separates operating-condition-driven nominal behavior from degradation-sensitive sensor deviations. This residual disentanglement mechanism, coupled with hybrid temporal modeling and hierarchical attention, enables robust and interpretable diagnostic inference across variable operating regimes.

Extensive experiments on NASA C-MAPSS, N-CMAPSS, and ALFA showed that PRIME consistently outperforms strong baseline models across the considered tasks. For Fault Detection (FD), the framework achieved near-ceiling performance, with accuracy and F_1 scores reaching approximately 98%–99% on the main benchmarks, while also reducing false alarm rates to below 4% on the most challenging datasets. For Fault Type Classification (FTC), PRIME reached macro-F_1 scores of 97.89% on FD003 and 96.93% on FD004, with corresponding macro-AUC values of 0.978 and 0.968, confirming strong multi-class discrimination under both single- and multi-regime settings.

For Health State Estimation (HSE), PRIME produced more faithful degradation trajectories and consistently reduced estimation error relative to all baselines, with RMSE ranging from 8.2 cycles on FD001 to 21.2 cycles on N-CMAPSS and R^2 values up to 0.956. When Remaining Useful Life (RUL) was projected from the learned health trajectory through the threshold-based mechanism introduced in this work, PRIME further achieved RMSE values between 10.6 and 22.1 cycles, together with substantial reductions in NASA score, such as 158 on FD001 and 318 on FD004. These results support the view that HSE provides a useful latent degradation representation for downstream prognostic assessment, even though RUL is not directly optimized by the model.

Beyond quantitative performance, PRIME provides built-in interpretability through residual analysis and hierarchical attention. The resulting sensor-level and temporal patterns are consistent with degradation-sensitive engine behavior, thereby improving transparency and practical trustworthiness in safety-critical PHM settings. In particular, the strongest gains were observed on the more challenging multi-regime benchmarks, suggesting that residual operating-condition disentanglement is especially beneficial when regime variability and degradation patterns are strongly entangled.

Overall, the contribution of PRIME lies not in introducing fundamentally new neural primitives, but in providing a principled integration of residual operating-condition disentanglement, hybrid temporal modeling, hierarchical interpretability, and multi-task diagnostic learning within a unified PHM framework. This integrated design offers a robust and scalable solution for aircraft engine diagnostics in complex operating environments.

The proposed residual disentanglement should therefore be understood as a practically effective approximation rather than as a universally valid physical decomposition. The present results support its usefulness on the considered benchmarks, but broader validation under unseen regimes, sensor drift, and more detailed uncertainty analysis will be necessary to further assess its deployment readiness. Similarly, while the observed gains are systematic and statistically supported, future work should better characterize the relationship between predictive improvements, architectural complexity, and cross-metric consistency.

Future work will focus on extending the framework toward better-calibrated prognostic projection, broader evaluation across heterogeneous fleet conditions, adaptive task balancing, and improved integration of complementary sensing modalities for real-world deployment.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conception and design, data curation, literature review, analysis and interpretation of results: Soukaina Mjahed and Ouail Mjahed; draft manuscript preparation: Ouail Mjahed, writing—review and editing, supervision: Soukaina Mjahed. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the conclusions of this study are freely available at the websites cited in references [1–3].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ALFA	AirLab Failure and Anomaly dataset
AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
C-MAPSS	Commercial Modular Aero-Propulsion System Simulation dataset
FAR	False Alarm Rate
FD	Fault Detection
FTC	Fault Type Classification
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HSE	Health State Estimation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
N-CMAPSS	Numerical Commercial Modular Aero-Propulsion System Simulation dataset
PHM	Prognostics and Health Management
PINN	Physics-Informed Neural Network
PRIME	Physics-guided Residual Integrated framework for Multi-task aircraft Engine diagnostics
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
RUL	Remaining Useful Life
t-SNE	t-distributed Stochastic Neighbor Embedding
TCN	Temporal Convolutional Network

References

1. Saxena A, Goebel K, Simon D, Eklund NHW. Damage propagation modeling for aircraft engine run-to-failure simulation. In: Proceedings of 2008 International Conference on Prognostics and Health Management; 2008 Oct 6–9; Denver, CO, USA. p. 1–9. doi:10.1109/PHM.2008.4711414.
2. Chatterjee S, Keprate A. Exploratory data analysis of the N-CMAPSS dataset for prognostics. In: Proceedings of 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM); 2021 Dec 13–16; Singapore. p. 1114–21. doi:10.1109/IEEM50564.2021.9673064.
3. Keipour A, Mousaei M, Scherer S. ALFA: a dataset for uav fault and anomaly detection. Int J Robot Res. 2021;20(2–3):515–20. doi:10.1177/0278364920966642.
4. Heimes FO. Recurrent neural networks for remaining useful life estimation. In: Proceedings of 2008 International Conference on Prognostics and Health Management; 2008 Oct 6–9; Denver, CO, USA. p. 1–6. doi:10.1109/PHM.2008.4711422.
5. Marcia L, Baptista ML, Henriques EMP, Prendinger H. Classification prognostics approaches in aviation. Measurement. 2021;182(5):109756. doi:10.1016/j.measurement.2021.109756.
6. Zheng S, Ristovski K, Farahat A, Gupta C. Long short-term memory network for remaining useful life estimation. In: Proceedings of 2017 IEEE International Conference on Prognostics and Health Management (ICPHM); 2017 Jun 19–21; Dallas, TX, USA. p. 88–95. doi:10.1109/ICPHM.2017.7998311.
7. Li X, Ding Q, Sun J-Q. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf. 2018;172(1–2):1–11. doi:10.1016/j.ress.2017.11.021.
8. Fang X, Xiao L, Shan Y. PBMT: a novel transformer-based model for accurate RUL prediction in industrial systems. In: Proceedings of 2024 Global Reliability and Prognostics and Health Management Conference (PHM-Beijing); 2024 Oct 11–13; Beijing, China. p. 1–8. doi:10.1109/PHM-Beijing63284.2024.10874751.
9. Kim G, Choi JG, Lim S. Using transformer and a reweighting technique to develop a remaining useful life estimation method for turbofan engines. Eng Appl Artif Intell. 2024;133(Pt E):108475. doi:10.1016/j.engappai.2024.108475.
10. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, et al. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng. 2017;29(10):2318–31. doi:10.1109/TKDE.2017.2720168.
11. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks. J Comput Phys. 2019;378:686–707. doi:10.1016/j.jcp.2018.10.045.
12. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3(6):422–40. doi:10.1038/s42254-021-00314-5.
13. Xiao D, Xiao H, Li R, Wang Z. Application of physical-structure-driven deep learning and compensation methods in aircraft engine health management. Eng Appl Artif Intell. 2024;136(Pt B):109024. doi:10.1016/j.engappai.2024.109024.
14. Fu S, Avdelidis NP, Plastropoulos A. Novel hybrid prognostics of aircraft systems. Electronics. 2025;14(11):2193. doi:10.3390/electronics14112193.
15. Chai A, Fang Z, Lian M, Huang P, Guo C, Yin W, et al. Hi-MDTCN: hierarchical multi-scale dilated temporal convolutional network for tool condition monitoring. Sensors. 2025;25(24):7603. doi:10.3390/s25247603.
16. Xu Z, Zhang Y, Miao Q. An attention-based multi-scale temporal convolutional network for remaining useful life prediction. Reliab Eng Syst Saf. 2024;250(3):110288. doi:10.1016/j.ress.2024.110288.
17. Wang Y, Wu M, Li X, Xie L, Chen Z. A survey on graph neural networks for remaining useful life prediction: methodologies, evaluation and future trends. Mech Syst Signal Process. 2025;229(1):112449. doi:10.1016/j.ymssp.2025.112449.
18. Li Z, Ma J, Fan R, Zhao Y, Ai J, Dong Y. Aircraft sensor fault diagnosis based on graphsage and attention mechanism. Sensors. 2025;25(3):809. doi:10.3390/s25030809.
19. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. Proc AAAI Conf Artif Intell. 2021;35(12):11106–15. doi:10.1609/aaai.v35i12.17325.

20. Zhong J, Li H, Chen Y, Huang C, Zhong S, Geng H. Remaining useful life prediction of rolling bearings based on ECA-CAE and autoformer. Biomimetics. 2024;9(1):40. doi:10.3390/biomimetics9010040.
21. Zhu J, Liang S, Ma Z, Huang X. Attention-based multi-modal learning for aircraft engine fan fault diagnosis. Aerosp Sci Technol. 2025;162(1):110194. doi:10.1016/j.ast.2025.110194.
22. Dinten R, Zorrilla M. Using time series foundation models for few-shot remaining useful life prediction of aircraft engines. Comput Model Eng Sci. 2025;144(1):239–65. doi:10.32604/cmes.2025.065461.
23. Szrama S. Adaptive cluster-count selection via deep Q-learning for turbofan engine prognostics and health monitoring. Neurocomputing. 2026;665:132294. doi:10.1016/j.neucom.2025.132294.
24. Szrama S. Turbofan engine health status prediction with heterogeneous ensemble deep neural networks. Int J Data Sci Anal. 2026;22(1):23. doi:10.1007/s41060-025-00989-4.
25. Duan Y, Xiao J, Li H, Zhang J. Cross-domain remaining useful life prediction based on adversarial training. Machines. 2022;10(6):438. doi:10.3390/machines10060438.
26. Ren L, Qin H, Cai N, Li B, Xie Z. A hybrid degradation evaluation model for aero-engines. Sustainability. 2023;15(1):29. doi:10.3390/su15010029.
27. Liu J, Yu Z, Zuo H, Fu R, Feng X. Multi-stageresidual life prediction of aero-engine based on real-time clustering and combined prediction model. Reliab Eng Syst Saf. 2022;225(11):108624. doi:10.1016/j.ress.2022.108624.
28. Willard J, Jia X, Xu S, Steinbach M, Kumar V. Integrating physics-based modeling with machine learning: a survey. arXiv:2003.04919. 2022.
29. Chao MA, Kulkarni C, Goebel K, Fink O. Fusing physics-based and deep learning models for prognostics. Reliab Eng Syst Saf. 2022;217(3):107961. doi:10.1016/j.ress.2021.107961.
30. Cummins L, Sommers A, Ramezani S, Mittal S, Jabour J, Seale M, et al. Explainable predictive maintenance: a survey of current methods, challenges and opportunities. IEEE Access. 2024;12(4):57574–602. doi:10.1109/ACCESS.2024.3391130.
31. He S, Wang S, Zhang R. A generalizable gated graph recurrent unit (Graph-GRU) network for nonlinear response prediction of cross-structures. Comput Struct. 2025;318(5):107968. doi:10.1016/j.compstruc.2025.107968.
32. Mienye ID, Swart TG, Obaido G. Recurrent neural networks: a comprehensive review of architectures, variants, and applications. Information. 2024;15(9):517. doi:10.3390/info15090517.
33. Wang M, Qin F. A TCN-linear hybrid model for chaotic time series forecasting. Entropy. 2024;26(6):467. doi:10.3390/e26060467.
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
35. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations, ICLR; 2015 May 7–9; San Diego, CA, USA. p. 1–15.
36. Saxena A, Goebel K. Turbofan engine degradation simulation data set. NASA Ames Progn Data Repos. 2008;18:878–87.
37. Ramasso E, Saxena A. Performance benchmarking and analysis of prognostic methods for C-MAPSS datasets. Int J Progn Health Manag. 2014;5(2):1–15. doi:10.36001/ijphm.2014.v5i2.2236.
38. Chao MA, Kulkarni C, Goebel K, Fink O. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. Data. 2021;6(1):5. doi:10.3390/data6010005.
39. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of 3rd International Conference for Learning Representations (ICLR); 2015 May 7–9; San Diego, CA, USA. p. 1–15.
40. Rajić V. Statistical hypothesis testing: a comprehensive review of theory, methods, and applications. Mathematics. 2026;14(2):300. doi:10.3390/math14020300.
41. Purwono I, Maarif A, Rahmaniar W, Imam H, Frisky AZK, ul Haq QM. Understanding of convolutional neural network (CNN): a review. Int J Robot Control Syst. 2023;2(4):739–48. doi:10.31763/ijrcs.v2i4.888.
42. Krichen M, Mihoub A. Long short-term memory networks: a comprehensive survey. Artif Intell. 2025;6(9):215. doi:10.3390/ai6090215.

43. Yang T, Cheng Y, Ren Y, Lou Y, Wei M, Xin H. A deep learning framework for sequence mining with bidirectional LSTM and multi-scale attention. In: Proceedings of 2nd International Conference on Innovation Management and Information System (ICIIS 2025); 2025 Apr 18–20; Shenzhen, China. p. 472–6. doi:10.1145/3745676.3745751.
44. Su L, Zuo X, Li R, Wang X, Zhao H, Huang B. A systematic review for transformer-based long-term series forecasting. Artif Intell Rev. 2025;58(3):80. doi:10.1007/s10462-024-11044-2.
45. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
46. Özcan H. Predictive maintenance in aircraft engine maintenance using the C-MAPSS dataset: performance comparison and evaluation of machine learning classification algorithms. Artif Intell Eng Des Anal Manuf. 2026;40:e4. doi:10.1017/S0890060426100249.
47. Sharma DB, Kodipalli A, Rao T, Rohini BR. Machine predictive maintenance classification using machine learning. In: Proceedings of 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA); 2023 Jun 22–23; Bengaluru, India. p. 308–13. doi:10.1109/CIISCA59740.2023.00066.
48. Al Hasib A, Rahman A, Khabir M, Shawon M. An interpretable systematic review of machine learning models for predictive maintenance of aircraft engine. arXiv:2309.13310. 2023.
49. Melkumian SA. Predictive maintenance analysis of turbofan engine sensor data. J Purdue Undergrad Res. 2024;14(1):8. doi:10.7771/2158-4052.1708.
50. Wu J, Kong L, Kang S, Zuo H, Yang Y, Cheng Z. Aircraft engine fault diagnosis model based on 1DCNN-BiLSTM with CBAM. Sensors. 2024;24(3):780. doi:10.3390/s24030780.
51. Li Y, Chen Y, Hu Z, Zhang H. Remaining useful life prediction of aero-engine enabled by fusing knowledge and deep learning models. Reliab Eng Syst Saf. 2023;229(4):108869. doi:10.1016/j.ress.2022.108869.
52. Zhu Q, Xiong Q, Yang Z, Yu Y. A novel feature-fusion-based end-to-end approach for remaining useful life prediction. J Intell Manuf. 2023;34(8):3495–505. doi:10.1007/s10845-022-02015-x.
53. Yang X, Gao X, Zheng H, Yang M, Liu Y. A hybrid prognosis method based on health indicator and wiener process: the case of multi-sensor monitored aero-engine. Eng Appl Artif Intell. 2025;144(Pt C):110099. doi:10.1016/j.engappai.2025.110099.
54. Liu CL, Xiao B, Hsu SS. Self-supervised learning for remaining useful life prediction using simple triplet networks. Adv Eng Inform. 2025;64(1):103038. doi:10.1016/j.aei.2024.103038.
55. Nunes P, Santos J, Rocha E. Combining generalized fault trees and k-LSTM ensembles for enhancing prognostics and health management. CIRP J Manuf Sci Technol. 2025;63:505–21. doi:10.1016/j.cirpj.2025.11.002.