I-GLIDE: Input Groups for Latent Health Indicators in Degradation Estimation

Lucas Thil[1,2] ✉, Jesse Read.[1], Rim Kaddah[2], and Guillaume Doquet[3]

[1] LIX Ecole Polytechnique, France thil@lix.polytechnique.fr
jesse.read@polytechnique.edu
[2] IRT SystemX, France lucas.thil@irt-system.fr rim.kaddah@irt-systemx.fr
[3] Safran Tech guillaume.doquet@safrangroup.com

Abstract. Accurate remaining useful life (RUL) prediction hinges on the quality of health indicators (HIs), yet existing methods often fail to disentangle complex degradation mechanisms in multi-sensor systems or quantify uncertainty in HI reliability. This paper introduces a novel framework for HI construction, advancing three key contributions. First, we adapt Reconstruction along Projected Pathways (RaPP) as a health indicator (HI) for RUL prediction for the first time, showing that it outperforms traditional reconstruction error metrics. Second, we show that augmenting RaPP-derived HIs with aleatoric and epistemic uncertainty quantification (UQ)—via Monte Carlo dropout and probabilistic latent spaces— significantly improves RUL-prediction robustness. Third, and most critically, we propose indicator groups, a paradigm that isolates sensor subsets to model system-specific degradations, giving rise to our novel method, I-GLIDE which enables interpretable, mechanism-specific diagnostics. Evaluated on data sourced from aerospace and manufacturing systems, our approach achieves marked improvements in accuracy and generalizability compared to state-of-the-art HI methods while providing actionable insights into system failure pathways. This work bridges the gap between anomaly detection and prognostics, offering a principled framework for uncertainty-aware degradation modeling in complex systems.

Keywords: Health Indicator · Latent Space · Degradation Modeling.

1 Introduction

Accurate RUL prediction is critical for enabling condition-based maintenance in complex engineering systems. A cornerstone of this task lies in deriving interpretable Health Indicators (HIs) that reliably capture subsystem degradation patterns. While autoencoder (AE)-based reconstruction errors have emerged as a popular HI-construction method, existing approaches suffer from two key limitations: (1) sensitivity to noise and epistemic uncertainty, which obscures degradation signals, and (2) a lack of granularity in disentangling subsystem-specific degradation behaviors. This work addresses these gaps by introducing a

novel Ensemble of Indicators framework, which advances traditional AE architectures through multi-head encoders and decoders designed to isolate degradation patterns across subsystems (e.g., fan, high-pressure compressor). We call this method I-GLIDE: Input Groups for Latent Health Indicators in Degradation Estimation.

Contrary to prior studies that treat UQ as an auxiliary feature, our proposed method I-GLIDE leverages this uncertainty to enhance HI robustness while maintaining explainability. We rigorously benchmark our approach against established latent-space HI methods—including Reconstruction along Projected Pathways (RaPP) [18, 30] and Monte Carlo (MC) dropout-based uncertainty estimation [9] demonstrating superior RUL prediction accuracy on the NASA C-MAPSS turbofan dataset [29] and the MILL NASA degradation dataset [2]. Our contributions are threefold:

1. **Systematic Analysis**: We identify and characterize critical limitations of existing AE derived HIs, notably their vulnerability to noise and inability to isolate subsystem-level degradation.
2. **Uncertainty-Aware Benchmarking**: By integrating aleatoric and epistemic UQ into latent-HI construction, we improve RUL estimation.
3. **I-GLIDE Framework**: We propose a multi-head AE architecture where each encoder-decoder pair targets distinct subsystems, enabling granular, explainable HI extraction, achieving state-of-the-art RUL prediction while providing insights into degradation mechanisms.

Table 1. Notation used in the paper.

Symbol	Description
σ_a, σ_e	Aleatoric, Epistemic uncertainty
$\mathcal{F}$	Function mapping HIs to a RUL
$\mathcal{X}$	Input data set with entries x
$\hat{x}$	reconstruction of input x, also noted as the target variable y
W_D	Decoder weight parameters
$g \in G$	Set of sub-complex systems indices g
z	Latent space of the AE
y	Target variable
$h_{g,l}$	hidden layer of group g h_g, at position l
$d_g(x)$	specific distance vector of groups $h_g(x) - h_g(\hat{x})$

2 Background and Related Works

2.1 RUL Prognostics and Health Indicators

Most industrial complex systems are built by the interdependencies of sub-complex systems; degradation in one component can propagate cascading effects,

triggering operational disruptions, escalating costs, safety risks, and—in extreme cases—catastrophic system-wide failures. As a result, the accurate RUL estimation of a complex system is heavily studied in engineering, particularly where costs and safety are associated. Early methodologies relied on stochastic approaches, such as threshold-based degradation signatures or empirical lifetime metrics (e.g., flight cycles or mileage) [8]. While methods prioritized identifying failure precursors or tracking cumulative usage they lacked adaptability to complex, non-linear degradation patterns. Timely and precise RUL prognostics not only curtails downtime and waste but also enables proactive maintenance, aligning operational decisions with evolving system health.

2.2 Evolution of HI Extraction

Early HI derivation prioritized interpretability through handcrafted statistical features (e.g., signal variance) or physics-based models. Using a a Bayesian framework enriched by expert knowledge to estimate failure probabilities, Lacaille [20] proposed a normalization pretreatment to derive standardized signatures interpretable as HIs by domain experts. However, such approaches depended heavily on predefined failure patterns and manual refinement, limiting their adaptability to heterogeneous operational conditions in non-stationary environments [6]. Hybrid approaches combined Kalman filters with NNs to model state-of-charge degradation [12], while others used neural networks (NNs) to learn a RUL representation to derive syncretic HIs [33]. Zhao et al. [35] used degradation pattern learning in the case of turbofan engines to predict the RUL. They extracted degradation patterns that helped characterize the nature of the degradation, which can itself be seen as a HI. Furthermore, their method was shown to improve the predictive capability of a NN towards RUL estimation.

AEs later emerged as a cornerstone method, using reconstruction errors from healthy-state training as implicit HIs [14, 24, 10]. Despite progress, these methods often assumed linear degradation trends or predefined failure modes, limiting adaptability to non-stationary systems. Other NN approaches later enabled data-driven prognostics, with Long Short-Term Memory (LSTM) architectures capturing temporal degradation in batteries [31] and turbofans [22]. However, early frameworks often bundled HI estimation with RUL prediction, risking conflated objectives where HIs were implicitly tuned to downstream tasks rather than intrinsic degradation patterns. This coupling became particularly evident in methods that embedded domain assumptions directly into HI design. For example, Jing et al. [16] incorporated exponential normalization of sensor data as an inductive bias in a Variational Autoencoder (VAE), aligning the HI with the CMAPSS dataset's predefined degradation trends. While this yielded robust RUL predictions, it effectively tied the HI to the target degradation profile, limiting adaptability to systems with non-exponential behaviors. Pillai and Vadakkepat [27] addressed this issue more directly, proposing a two-stage architecture that decoupled HI feature discovery from RUL regression. Their approach improved generalizability by isolating degradation modeling from task-specific optimization, though challenges persisted in interpretability and subsystem-specific analysis.

Latent-Space HI Refinement More recent advances focused on refining HI quality through latent-space analysis. Kim et al. [18] introduced the RaPP method, projecting latent representations from an AE's encoder to compute distance metrics that outperformed classical reconstruction errors. González-Muñiz et al. [11] validated this paradigm shift, demonstrating that latent-space metrics from RaPP consistently surpass input-space approaches in HI quality. Their work highlighted the latent space between encoder and decoder as a rich source of degradation signals, though subsystem-specific trends remained obscured by holistic aggregation. Despite these innovations, mapping HIs to RUL remains fraught with challenges. Many approaches employ black-box models or simplistic linear mappings [23], neglecting context-dependent HI interpretations under varying operational conditions. For instance, a high reconstruction error might indicate severe degradation in one context but sensor noise in another—a nuance often lost in end-to-end frameworks. Recent benchmarking by Rombach et al. [28] underscores this gap, advocating for feature engineering to improve HI interpretability while maintaining correlation with ground truth degradation.

2.3 Uncertainty-Aware Subsystem Modeling

Uncertainty quantification in NNs can be achieved through MC dropout [1]. Variational AEs (VAEs) [3], can disentangle aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model ambiguity) to isolate distinct sources of unpredictability. While probabilistic frameworks optimize maintenance via confidence intervals [25], UQ is often treated as a post-hoc refinement rather than a core HI component. Deterministic AEs, for instance, cannot isolate aleatoric uncertainty due to fixed latent spaces—a limitation addressed by variational architectures [32]. Ensemble methods further reduce uncertainty [34], yet their application to subsystem-aware HIs remains underexplored.

2.4 Monotonicity and Degradation Dynamics

RUL is typically modeled as a monotonic function of the State of Health (SOH), declining from 100% (pristine) to 0% (failure). While mechanical wear rarely reverses, subsystem interactions (e.g., turbine degradation accelerating fan wear) introduce non-stationary dynamics [4]. This necessitates HIs that isolate localized degradation while preserving system-wide coherence—a gap addressed by our subsystem-aware architecture. Similarly, we model RUL estimation as a function $\mathcal{F}$ of the HIs, mapping their values to the corresponding RUL.

3 Proposed Approach: I-GLIDE

In order to produce better HIs, we build on foundational assumptions about degradation dynamics and their statistical relationships to RUL established earlier, and we begin by formalizing our UQ for prognostic tasks. We then introduce a novel architecture that disentangles subsystem-specific degradation signatures

by managing sensor groups and operational variabilities at the component level, while adapting the RaPP method to mitigate cross-component interference. Finally, we propose a data-driven strategy to validate constructed HIs through direct RUL estimation, demonstrating their prognostic utility. Each phase is rigorously evaluated via empirical case studies (Section 4), ensuring robustness across diverse degradation scenarios. Our notation is summarized in Table 1.

3.1 Uncertainty Quantification

Uncertainty in prognostics arises from two primary sources: aleatoric (σ_a), inherent to data noise and irreducible even with additional observations, and epistemic (σ_e), stemming from model limitations and reducible through improved architectures or training [15, 17].

In order to produce high-quality HIs we make use of the UQ capabilities of AE architectures with an underlying change. Our epistemic UQ focuses on the scalar reconstruction error $\epsilon = \|x - \hat{x}\|_2$ instead of the full-dimensional decoder output $\hat{x} = y$. This aligns with prognostics frameworks where ϵ serves as a health indicator (HI), reducing dimensionality for easier integration with downstream RUL prediction models (e.g., $\mathcal{F}$). We avoided using raw y (the reconstruction) as a standalone HI directly because it is 1) outperformed by RaPP methods [18, 11] and 2) our RUL predictor model $\mathcal{F}$ showed high variance in selecting the best variables when both RaPP and y were fed as inputs. Thus we dropped y as a HI and instead focused in introducing MC dropout to quantify ϵ uncertainties as a HI which showed to be a better complement. Therefore, the disentangled UQ is performed through the aggregation of our $\epsilon_1..\epsilon_n$ over n MC samples in our VAE architecture:

$$\sigma_a = \mathrm{Var}\left(\epsilon_1, \ldots, \epsilon_n | \text{ fixed } \mathrm{W_D}\right), \quad \sigma_e = \mathrm{Var}\left(\epsilon_1, \ldots, \epsilon_n | \text{ fixed } z\right).$$

As z is deterministic, aleatoric uncertainty σ_a cannot be isolated in AEs, rendering it undefined. Thus, we can only compute σ_e in the case of a vanilla AE. This highlights the advantage of VAEs for joint uncertainty estimation.

3.2 Architecture

Building on the above foundations, our proposed architecture extends the traditional AE and VAE frameworks by introducing multiple encoder-decoder pairs for each sensor group, which are then integrated through a shared latent space, as illustrated in Fig. 1 (3.2). This design addresses the non-stationarity of sensor signals by disentangling subsystem-specific degradation dynamics in the latent space. This separation allows us to apply the RaPP [18] method individually to each encoder. By projecting the activations of the hidden spaces $h_g(x)$ corresponding to isolated sensor groups $g \in G$, we aim to achieve more comprehensive feature extraction, enabling the construction of specific health indicators (HIs).

We derive this architecture with two different latent spaces: the first one being in the way of traditional AE named I-GLIDE$_{\mathrm{AE}}$, and in the second version

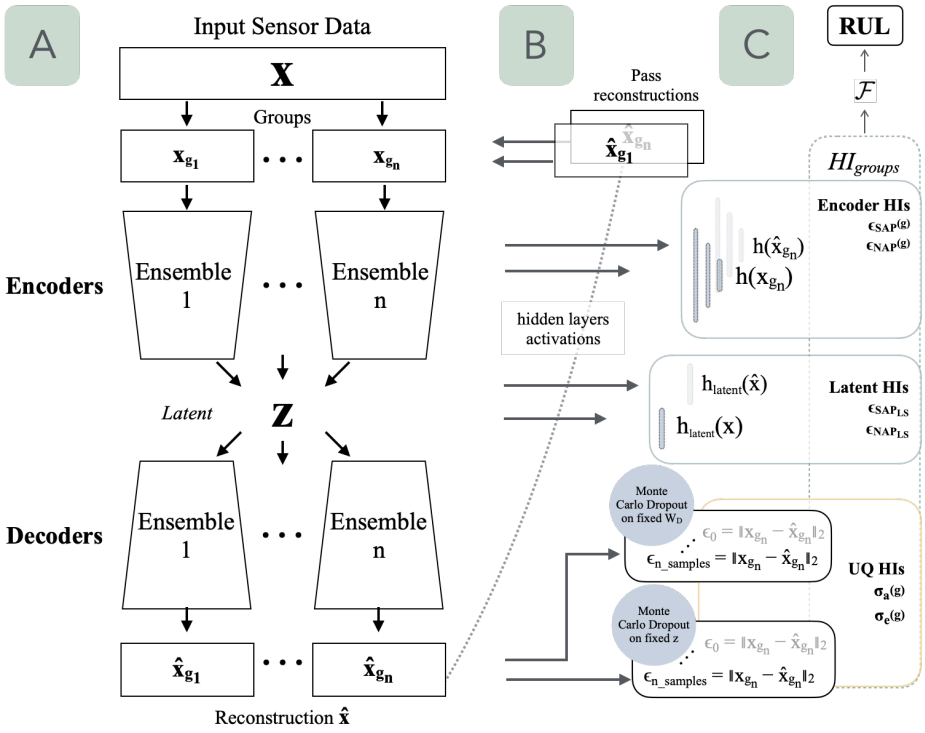


Fig. 1. I-GLIDE Architecture Framework – **A:** Subsystem-specific encoder-decoder heads learn distinct latent representations, fused into a shared latent space z via reconstruction loss (trained on healthy data). **B:** HIs are extracted using RaPP metrics [11] and UQ [19] over full trajectories. **C:** Aggregated HIs are used to predict RUL, trained via a Random Forest (RF) regressor $\mathcal{F}$.

where the latent is a Gaussian type distribution in the manner of VAEs named I-GLIDE$_{\text{VAE}}$. In the latter, we can leverage the variational inference aspect of the architecture.

3.3 Adapting Domain-Specific Latent Space Health Indicators

A main novelty of I-GLIDE is to adapt RaPP [18], traditionally used in monolithic architectures, for each subsystem in our multi-autoencoder framework. By moving away from the monolithic approach, I-GLIDE computes group-specific health indicators (HIs) for each sensor group g. Unlike the original RaPP framework, which operates on a single encoder-decoder pair, our architecture independently calculates $\epsilon_{\text{SAP}^{(g)}}$ and $\epsilon_{\text{NAP}^{(g)}}$ for each group g, leveraging dedicated encoders and decoders per subsystem. These components share a cohesive latent space z, preserving global system coherence while isolating localized anomalies. For a sensor group g, let $h_{g,l}(x) \in \mathbb{R}^{n_l}$ denote the activations of the l-th en-

coder layer (with n_l-dimensional output) for input x, and $h_{g,l}(\hat{x}_g)$ represent the reconstructed activations across L layers ($l = 1, \ldots, L$).

We redefine two HIs per sensor group:

$$\epsilon_{\text{SAP}^{(g)}}(x) = \|h_g(x) - h_g(\hat{x})\|_2 \tag{1}$$

which computes the first RaPP metric: the Simple Aggregation along Pathway (SAP) [18] as the Euclidean distance between original and reconstructed activations across all layers l. For the second RaPP metric, Normalized Aggregation along Pathway (NAP), we first derive the group-specific distance vector $d_g(x) = h_g(x) - h_g(\hat{x})$, where $h_g(x) = [h_{g,1}(x), \ldots, h_{g,L}(x)]$ concatenates activations across all layers l. Given a training set $\mathcal{X}$, let D_g be a matrix whose rows correspond to $d_g(x_i)$ for $x_i \in \mathcal{X}$, and let $\bar{D}_g$ denote the column-wise centered version of D_g. The NAP metric for group g is then:

$$\epsilon_{\text{NAP}^{(g)}}(x) = \left\| (d_g(x) - \mu_{\mathcal{X}})^\top V_g \Sigma_g^{-1} \right\|_2 . \tag{2}$$

Here, $\mu_{\mathcal{X}_g} \in \mathbb{R}^{n_l \cdot L}$ is the column-wise mean of D_g, $\Sigma_g \in \mathbb{R}^{k \times k}$ is a diagonal matrix containing the singular values of $\bar{D}_g$, and $V_g \in \mathbb{R}^{(n_l \cdot L) \times k}$ contains the right singular vectors from the singular value decomposition (SVD) of $\bar{D}_g$, with k denoting the rank of $\bar{D}_g$.

This design allows the model to isolate sensor group contributions, where anomalies in specific sensor groups are preserved without being diluted by nominal signals from other groups and this results in enhanced interpretability as HIs directly map to physical sensor groups, aiding root-cause analysis.

Building on González et al. [11], where latent-space RaPP metrics outperform encoder-derived counterparts for HI construction, our method integrates both approaches. We compute latent-space z metrics ($\epsilon_{\text{NAP}_{\text{LS}}}$, $\epsilon_{\text{SAP}_{\text{LS}}}$), with $\epsilon_{\text{SAP}_{\text{LS}}}$ derived from all the data and not by individual groups.

3.4 Final set of HIs

The full set of HIs produced by I-GLIDE are then aggregated with our UQ as the set $\text{HI}_{\text{groups}} = \{\epsilon_{\text{SAP}^{(g)}}, \epsilon_{\text{NAP}^{(g)}}, \epsilon_{\text{SAP}_{\mathbf{LS}}}, \epsilon_{\text{NAP}_{\mathbf{LS}}}, \sigma_{a^{(g)}}, \sigma_{e^{(g)}}\} \forall g \in G$ where $\sigma_{a^{(g)}}, \sigma_{e^{(g)}}$ are respectively the aleatoric and espitemic uncertainties computed for each group g. We also compare with the monilithic architecture where the inputs x are not divided into subgroups, and thus our set of HIs is defined as $\text{HI}_{\text{mono}} = \{\epsilon_{\text{SAP}}, \epsilon_{\text{NAP}}, \epsilon_{\text{SAP}_{\mathbf{LS}}}, \epsilon_{\text{NAP}_{\mathbf{LS}}}, \sigma_a, \sigma_e\}$.

To evaluate their predictive capabilities, we train a meta regressor $\mathcal{F}(.)$ on the task of RUL estimation. We also compare with the previous set of RaPP indicators from Gonzàlez by define $\text{HI}_{\text{González}} = \{\epsilon_{\text{NAP}_{\text{LS}}}, \epsilon_{\text{SAP}_{\text{LS}}}\}$ [11].

4 Experiments

We aim to find out whether augmenting latent space HIs (RaPP metrics) with UQ, contributes to better understanding of degradation mechanisms in complex

Table 2. The four C-MAPSS dataset subsets and their description, with associated number of operating conditions and amount of degradation fault modes origins.

	FD001	**FD002**	**FD003**	**FD004**
Train trajectories	100	260	100	248
Test trajectories	100	259	100	249
Conditions	1	6	1	6
Fault modes	1	1	2	2

system in the perspective of RUL estimation. In a second step, we'd like to test whether introducing an architecture able to disentangle sub-systems degradation mechanisms can further improve this RUL estimation, through better understanding of the system from data, and minimal domain-knowledge. Our code is available at: https://github.com/LucasStill/I-GLIDE for reproduction purposes.

4.1 Datasets

We evaluate our framework on two datasets. The C-MAPSS dataset [29] a benchmark for degradation modeling, contains simulated run-to-failure trajectories of jet engines generated using NASA's C-MAPSS simulator. Each multivariate time series corresponds to a unique engine operating under varying conditions, divided into four subsets (FD001–FD004; see Table 2). During training, we focus on samples with RUL ≤ 80 timesteps ($R_{early} = 80$) to prioritize early degradation signals while retaining healthy-state representations. For testing, we follow the established protocol by prior works with $R_{early} = 125$ to enable direct comparison. The test set contains truncated trajectories that stop before the point of failure, and the task is to predict this last available value on the trajectory. We present the partitioning of the different groups in table 3.

To further validate our approach and compare the effects of subsystem group separation, we test on the MILL NASA dataset, which records 167 unique tool wear progression during milling experiments under varied conditions (depth of cut, feed rate, material). Sensor signals (acoustic, vibration, current) track wear, with failure defined at wear=0.70 (initial wear=0). We classify samples as healthy (wear ≤ 0.20) during training and evaluate degradation over three test phases: complete trajectories, moderate degradation (wear > 0.20), and severe degradation (wear > 0.50). Each sensor contains a total of 9000 entries, but some have missing data which we fill through interpolation with neighboring values.

For both datasets, we will apply our method to create our sets of HIs and measure their predictive capabilities over RUL prediction, using RMSE metric which is commonly used on these datasets [7].

4.2 Experimental Methodology

To validate our framework, we adopt a prognostics-centric evaluation protocol that directly benchmarks HIs by their ability to predict RUL—the ultimate

Table 3. Grouping of sensors in the CMAPS dataset.

Group	Sensor ID	Description
Fan	s_1	Total temperature at fan inlet (°R)
	s_5	Pressure at fan inlet (psia)
	s_8	Physical fan speed (rpm)
	s_13	Corrected fan speed (rpm)
	s_18	Demanded fan speed (rpm)
	s_19	Demanded corrected fan speed (rpm)
LPC	s_2	Total temperature at LPC outlet (°R)
HPC	s_3	Total temperature at HPC outlet (°R)
	s_7	Total pressure at HPC outlet (psia)
	s_11	Static pressure at HPC outlet (psia)
Core	s_9	Physical core speed (rpm)
	s_14	Corrected core speed (rpm)
Pressure Turbine	s_4	Total temperature at LPT outlet (°R)
	s_20	HPT coolant bleed (lbm/s)
	s_21	LPT coolant bleed (lbm/s)
Other	s_6	Total pressure in bypass-duct (psia)
	s_10	Engine pressure ratio (P50/P2) (-)
	s_12	Ratio of fuel flow to Ps30 (pps/psia)
	s_15	Bypass Ratio (-)
	s_16	Burner fuel-air ratio (-)
	s_17	Bleed Enthalpy (-)

objective of HI construction. We first compare the RaPP-based HIs proposed by González et al. ($HI_{González}$) [11] against a enhanced variants: HI_{mono}, which integrates encoder-level RaPP metrics [18], with UQ. We use a single timestep for each created HI. Critically, we bypass classical HI metrics like monotonicity or trendability, which often fail to correlate with actionable prognostic value, and instead train a random forest (RF) regressor $\mathcal{F}$ to map HIs to RUL. This choice reflects a key design principle: HI quality should be first judged by its downstream utility in prognostics.

We then introduce our I-GLIDE architecture, instantiated as I-GLIDE$_{AE}$ and I-GLIDE$_{VAE}$, which generates subsystem-specific HIs (HI_{groups}) by isolating sensor-group degradation patterns. These are benchmarked against monolithic AE/VAE counterparts under identical RF training protocols, ensuring fair comparison. By using a simple, non-temporal model like RF, we deliberately decouple HI quality from algorithmic sophistication, isolating how architectural choices (monolithic vs. subsystem-specific) impact prognostics performance.

4.3 Results

When comparing the RUL estimation capabilities of $HI_{González}$ and HI_{mono} as shown in table 4, we find that HI_{mono} consistently outperforms $HI_{González}$ across all C-MAPSS subsets (FD001-FD004). For example, HI_{mono} derived from AEs reduces RMSE by 22.95% on average compared to $HI_{González}$, with particularly notable gain on FD002 (15.71 vs 22.91) and FD003 (8.07 vs 12.03). Similar trends hold for VAEs, where HI_{mono} achieves a 28.44% average RMSE improvement, underscoring the value of UQ in stabilizing HI quality. The same can be observed for the MILL dataset in table 5. This ablation study demonstrates that a broader coverage of latent HIs with UQ, collectively strenghtens RUL predictive capabilities, even before subsytem-specific modeling.

Next, we deploy I-GLIDE, which explicitly disentangles subsytem degradation (e.g., HPC, fan, turbine in C-MAPSS) by grouping sensor signals into functionally coherent components (exacts group choices are presented in the appendix). Compared to monolithic architectures, I-GLIDE achieves superior robustness, as evidenced by its 39.96% reduction of standard deviation in RMSE across C-MAPSS subsets from AE-based HIs (6). For VAEs, gains are even more pronounced: I-GLIDE_{VAE} reduces RMSE by 39.03% and standard deviation by 56.07%, resolving the instability seen in monolithic VAEs (e.g., FD002/FD003 variance). This subsystem isolation proves critical on FD004-the most complex C-MAPSS subset-where I-GLIDE's average results set a new state-of-the-art performance with a RMSE of 14.19 despite using only a RF regressor for RUL prediction.

On the MILL dataset, I-GLIDE's subsytem-specific HIs improve RUL prediction across all degradation phases (healthy, moderate, severe), with I-GLIDE_{AE}-driven HIs achieving the lowest RMSE in every scenario. VAE gains are subtler, likely due to MILL's lower inherent complexity, or high dimensional space which limits the benefits of variational inference.

Table 4. Comparison of sets of HIs extracted from different architectures to predict the RUL RMSE on C-MAPSS test dataset using a Random Forest for $\mathcal{F}$. Best models shown.

HI Extractor	**HI Set for $\mathcal{F}(.)$**	**FD001**	**FD002**	**FD003**	**FD004**	**Avg.**
AE	$HI_{González}$ [11]	11.43	22.91	12.03	16.78	15.79
	HI_{mono}	10.53	**15.71**	**8.07**	14.35	12.17
VAE	$HI_{González}$ [11]	27.56	28.62	24.36	22.33	25.72
	HI_{mono}	18.77	19.44	15.59	19.81	18.40
I-GLIDE_{AE}	HI_{groups}	**9.47**	16.18	8.29	12.32	11.57
I-GLIDE_{VAE}	HI_{groups}	12.33	16.76	8.5	**11.4**	12.25

Table 5. RUL MILL Dataset Benchmark on three wear levels. I-GLIDE HIs consistently outperform the monolithic counterpart in RMSE for RUL prediction [2].

Model Name, HI set for RF	**Wear 0.0-0.70**	**Wear 0.20-0.70**	**Wear 0.50-0.70**
AE, $HI_{González}$ [11]	23.78	24.34	22.33
AE, HI_{mono}	16.14	16.47	16.25
I-GLIDE$_{AE}$, HI_{groups}	**13.64**	**14.37**	**16.17**
VAE, $HI_{González}$ [11]	27.84	27.92	27.14
VAE, HI_{mono}	22.32	22.46	23.47
I-GLIDE$_{VAE}$, HI_{groups}	**21.76**	**22.29**	**23.13**

Table 6. Average model performances across 10 runs over C-MAPSS subsets using RMSE (mean ± standard deviation). Bold: best results per subset; underline: outperforms methods without HIs. Last column provides average improvement over the previous row.

Model, HI Set	**FD001**	**FD002**	**FD003**	**FD004**	**Avg.**	**Improvement**
AE, $HI_{González}$ [11]	19.00 ±4.78	25.69 ±4.19	18.38 ±6.18	19.46 ±2.46	20.63 ±4.40	–
AE, HI_{mono}	13.14 ±2.50	20.35 ±3.46	13.87 ±5.07	17.73 ±3.56	16.27 ±3.65	+21.13% +17.15%
I-GLIDE$_{AE}$, HI_{groups}	**12.11** ±2.72	22.01 ±2.88	**10.23** ±1.85	<u>14.92</u> ±1.31	14.82 ±2.19	+8.94% +39.96%
VAE, $HI_{González}$ [11]	34.13 ±3.71	31.05 ±1.89	27.25 ±2.58	25.23 ±2.03	29.42 ±2.55	–
VAE, HI_{mono}	27.19 ±5.97	22.81 ±2.86	24.64 ±5.26	22.89 ±1.82	24.38 ±3.98	+17.10% -55.83%
I-GLIDE$_{VAE}$, HI_{groups}	15.32 ±2.08	**18.83** ±1.51	11.12 ±2.29	**<u>14.19</u>** ±1.11	14.87 ±1.75	+39.03% +56.07%

5 Discussion

Remarkably, when looking at the best produced models, even with a RF-a model far simpler than deep learning baselines-our HIs match or exceed prior SOTA on three out of four C-MAPSS benchmarks as shown in table 7. This paradox highlights that HI quality, not model complexity, drives prognostics success. When looking at the expected accuracies of the different models, we see that I-GLIDE has lower standard deviations, being more robust to prediction, which explains why in two cases the best model was a monolithic AE: despite showing great performance on a single set, its high standard deviation shown in table 6 indicates it would not be robust on a broader test set, or in real-life conditions. This is why I-GLIDE offers solid perspectives towards more robust predictions.

Monolithic AEs struggle to disentangle subsystem-specific degradation, which we illustrate with a HI plot of the trajectories in Figure 5. For Engine 1 (FD001), where HPC degradation is the source, HI_{mono} shows weak latent-space (z) sensitivity to subsystem dynamics. This occurs because deeper layers in monolithic AEs compress sensor signals into a global representation, obscuring non-

Table 7. Comparison of I-GLIDE method for HI extraction benchmarked to predict a RUL, compared with best known approaches. In bold are the best results for each subset. Most previous methods were predicting a RUL from transformed sensor data without producing HIs, contrarily to our method which does provide HIs.

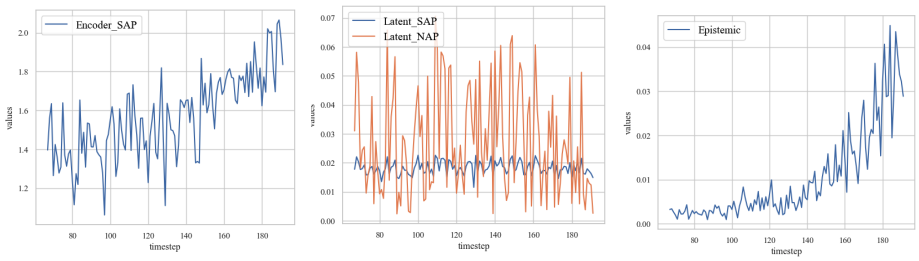
Model	**FD001**	**FD002**	**FD003**	**FD004**
MLP [36]	37.56	80.03	37.39	77.37
CNN [36]	18.45	30.29	19.82	29.16
CNN-LSTM [26]	11.17	-	9.99	-
MS-DCNN [21]	11.44	19.35	11.67	22.22
VAE + RNN [5]	11.44	24.12	14.88	26.54
MLE(4X)+CCF [27]	11.57	18.84	11.83	20.78
RVE [5]	13.42	14.92	12.51	16.37
Probabilistic RUL CNN [7]	12.42	**13.72**	12.16	15.95
I-GLIDE$_{\text{AE}}$ + RF (ours)	**9.47**	16.18	**8.29**	12.32
I-GLIDE$_{\text{VAE}}$ + RF (ours)	12.33	16.76	8.5	**11.4**

stationary interactions (e.g., HPC wear indirectly altering turbine behavior). In contrast, Figure 5 reveals how I-GLIDE isolates these dynamics: the HPC encoder HI exhibits a clear upward trend, while the turbine HI shifts abruptly as degradation propagates—a causal linkage masked in monolithic architectures. Notably, the shared latent z in I-GLIDE still captures the composite degradation trend, and subsystem-specific decoders also localize fault origins (e.g., rising epistemic uncertainty in HPC vs. stable turbine estimates). This explains why HI_{mono} underperforms—it conflates cross-subsystem effects into a single noisy signal, while our I-GLIDE overcomes these restrains. In future work, we would like to formalize methods to interpret such causal relationships between HIs, identify noise patterns in the degradation signals, and apply it to maintenance tasks.

Traditional HI metrics (monotonicity, trendability, prognosability) often produce misleading scores (e.g., near-perfect prognosability) that poorly correlate with actual RUL prediction. Worse, they ignore subsystem-specific degradation, obscuring actionable insights. Our framework addresses this by directly linking HI quality to RUL prediction accuracy—a metric aligned with real-world decision-making. By disentangling subsystem trends (e.g., turbine wear vs. fan imbalance), I-GLIDE enables targeted fault diagnosis and maintenance planning.

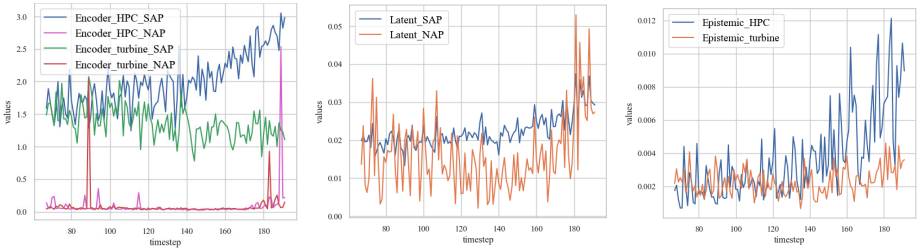
While our framework advances subsystem-aware HIs, several constraints merit consideration. First, it is worth noting that both the C-MAPSS and MILL datasets model exponential degradation patterns, which oversimplify real-world scenarios where industrial systems often exhibit linear or piecewise degradation trends. Real-world applications also introduce complex noise profiles (e.g., cyclic sensor artifacts) and heterogeneous failure modes that our method may not optimally capture without tailored adaptations.

Readers should be aware that our architecture assumes strictly monotonous degradation, limiting its ability to model recovery phases—a critical shortcoming for systems where transient improvements occur, such as medical devices



(a) Latent Encoder HI (b) Latent z HIs (c) Epistemic UQ HI

Fig. 2. AE HI trajectories for Engine 1 for the monolithic architecture. We can observe that the HIs model a degradation, but cannot distinguish sub-system components. We only show the SAP metric for the encoder HIs because NAP shows extreme values.



(a) Latent Encoder HIs (b) Latent z HIs (c) Epistemic UQ HI

Fig. 3. I-GLIDE$_{\text{AE}}$ HI trajectories for Engine 1, comparing degradation effects on HPC and Turbine. Latent encoder HIs (a) show rising HPC degradation and reduced Turbine HIs due to cross-component effects. System-wide latent z HIs trends are in (b). Epistemic uncertainty (c) rises sharply for HPC as degradation progresses, remaining stable for the Turbine until late-cycle HPC interference. UQ confirms causal cross-component effects without confusing intrinsic health states.

supporting patient recovery or aircraft exiting high-stress environments. Furthermore, our subsystem groupings rely on domain heuristics; while this aligns with prior work, poorly defined sensor groupings could propagate biases into the latent representations, undermining HI interpretability.

6 Conclusion and Future Work

This work establishes the first prognostics benchmark for evaluating Health Indicators (HIs) generated via the RaPP methods, demonstrating that integrating uncertainty quantification significantly enhances their predictive capabilities. Building on this foundation, we introduce I-GLIDE, a novel framework that learns subsystem-specific latent representations through dedicated encoder-decoder pairs. By isolating degradation mechanisms (e.g., HPC degradation vs.

turbine wear) while maintaining global system dynamics via a shared latent space, I-GLIDE captures nuanced failure modes without compromising system-level coherence. The resulting high-quality HIs achieve state-of-the-art performance on the C-MAPSS dataset, surpassing existing deep learning benchmarks using only a simple Random Forest regressor.

Our subsystem-specific HIs advance prognostics but invite refinement. Temporal improvements—like extending observation windows—could better resolve slow degradation signatures and transient noise, aligning HI trajectories with real-world failure timelines. Coupling uncertainty-specific t-SNE visualizations with expert annotations could map latent clusters to physical degradation stages, bridging data-driven insights with domain knowledge.

A promising direction involves modeling causal subsystem interactions via architectures like graph neural networks, trained on fused HIs to disentangle degradation propagation (e.g., turbine-to-compressor wear). This would scale prognostics to systems with complex interdependencies.

Critically, our results show that high-quality HIs paired with simple models (e.g., RF) outperform deep learning on raw data—a "data is gold" paradigm. Future efforts should prioritize refining physics-aware HI representations—grounded in subsystem dynamics and enriched with UQ to unlock generalizable, trustworthy RUL prediction across grounded industrial domains.

Acknowledgments. This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute within the JNI3 project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U., Makarenkov, V. & Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion. 76 pp. 243-297 (2021,12), https://linkinghub.elsevier.com/retrieve/pii/S1566253521001081
2. Agogino, A. & Goebel, K. Milling Data Set. NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA. (2007)
3. Akbari, A. & Jafari, R. Personalizing Activity Recognition Models Through Quantifying Different Types of Uncertainty Using Wearable Sensors. IEEE Transactions On Biomedical Engineering. 67, 2530-2541 (2020,9), https://ieeexplore.ieee.org/document/8949726/
4. Amini, A., Soleimany, A., Karaman, S. & Rus, D. Spatial Uncertainty Sampling for End-to-End Control. (arXiv,2018), https://arxiv.org/abs/1805.04829
5. Costa, N. & Sánchez, L. Variational encoding approach for interpretable assessment of remaining useful life estimation. Reliability Engineering & System Safety. 222 pp. 108353 (2022,6), https://linkinghub.elsevier.com/retrieve/pii/S0951832022000321

6. Côme, E. Aircraft engine health monitoring using Self-Organizing Maps.. 10th Industrial Conference, ICDM 2010, Berlin, Germany. pp. pp.405-417
7. De Pater, I. & Mitici, M. Novel Metrics to Evaluate Probabilistic Remaining Useful Life Prognostics with Applications to Turbofan Engines. PHM Society European Conference. 7, 96-109 (2022,6), https://papers.phmsociety.org/index.php/phme/article/view/3320
8. Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W. & Ducoffe, M. Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications Of Artificial Intelligence. 92 pp. 103678 (2020,6), https://linkinghub.elsevier.com/retrieve/pii/S0952197620301184
9. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. (2016), https://arxiv.org/abs/1506.02142
10. Gherbi, E., Hanczar, B., Janodet, J. & Klaudel, W. An Encoding Adversarial Network for Anomaly Detection. Proceedings Of The Eleventh Asian Conference On Machine Learning. 101 pp. 188-203 (2019,11,17), https://proceedings.mlr.press/v101/gherbi19a.html
11. González-Muñiz, A., Díaz, I., Cuadrado, A. & García-Pérez, D. Health indicator for machine condition monitoring built in the latent space of a deep autoencoder. Reliability Engineering & System Safety. 224 pp. 108482 (2022,8)
12. He, W., Williard, N., Chen, C. & Pecht, M. State of charge estimation for Li-ion batteries using neural network modeling and unscented Kalman filter-based error cancellation. International Journal Of Electrical Power & Energy Systems. 62 pp. 783-791 (2014,11), https://linkinghub.elsevier.com/retrieve/pii/S0142061514002646
13. Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. (arXiv,2012,7), http://arxiv.org/abs/1207.0580, arXiv:1207.0580 [cs]
14. Huang, L., Pan, X., Liu, Y. & Gong, L. An Unsupervised Machine Learning Approach for Monitoring Data Fusion and Health Indicator Construction. Sensors. 23, 7239 (2023,8), https://www.mdpi.com/1424-8220/23/16/7239
15. Hüllermeier, E. & Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. (arXiv,2019), https://arxiv.org/abs/1910.09457
16. Jing, T., Zheng, P., Xia, L. & Liu, T. Transformer-based hierarchical latent space VAE for interpretable remaining useful life prediction. Advanced Engineering Informatics. 54 pp. 101781 (2022,10), https://linkinghub.elsevier.com/retrieve/pii/S1474034622002397
17. Kendall, A. & Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. (arXiv,2017), https://arxiv.org/abs/1703.04977
18. Ki Hyun, K., Sangwoo, S. & Yongsub, L. RaPP: Novelty Detection with Reconstruction along Projection Pathway.
19. Kingma, D. & Welling, M. Auto-Encoding Variational Bayes. (arXiv,2013), https://arxiv.org/abs/1312.6114
20. Lacaille, J. Standardized failure signature for a turbofan engine. 2009 IEEE Aerospace Conference. pp. 1-8 (2009,3), http://ieeexplore.ieee.org/document/4839670/
21. Li, H., Zhao, W., Zhang, Y. & Zio, E. Remaining useful life prediction using multi-scale deep convolutional neural network. Applied Soft Computing. 89 pp. 106113 (2020,4), https://linkinghub.elsevier.com/retrieve/pii/S1568494620300533

22. Li, X., Ding, Q. & Sun, J. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety. 172 pp. 1-11 (2018,4), https://linkinghub.elsevier.com/retrieve/pii/S0951832017307779
23. Liu, D., Zhou, J., Liao, H., Peng, Y. & Peng, X. A Health Indicator Extraction and Optimization Framework for Lithium-Ion Battery Degradation Modeling and Prognostics. IEEE Transactions On Systems, Man, And Cybernetics: Systems. 45, 915-928 (2015,6), http://ieeexplore.ieee.org/document/7018028/
24. Martinelli, M., Tronci, E., Dipoppa, G. & Balducelli, C. Electric Power System Anomaly Detection Using Neural Networks. Knowledge-Based Intelligent Information And Engineering Systems. 3213 pp. 1242-1248 (2004), http://link.springer.com/10.1007/9783540301325168
25. Mitici, M., De Pater, I., Barros, A. & Zeng, Z. Dynamic predictive maintenance for multiple components using data-driven probabilistic RUL prognostics: The case of turbofan engines. Reliability Engineering & System Safety. 234 pp. 109199 (2023,6), https://linkinghub.elsevier.com/retrieve/pii/S095183202300114X
26. Peng, C., Chen, Y., Chen, Q., Tang, Z., Li, L. & Gui, W. A Remaining Useful Life Prognosis of Turbofan Engine Using Temporal and Spatial Feature Fusion. Sensors. 21, 418 (2021,1), https://www.mdpi.com/1424-8220/21/2/418
27. Pillai, S. & Vadakkepat, P. Two stage deep learning for prognostics using multi-loss encoder and convolutional composite features. Expert Systems With Applications. 171 pp. 114569 (2021,6), https://linkinghub.elsevier.com/retrieve/pii/S0957417421000105
28. Rombach, K., Michau, G., Bürzle, W., Koller, S. & Fink, O. Learning Informative Health Indicators Through Unsupervised Contrastive Learning. IEEE Transactions On Reliability. pp. 1-13 (2024), https://ieeexplore.ieee.org/document/10531793/
29. Saxena, A., Goebel, K., Simon, D. & Eklund, N. Damage propagation modeling for aircraft engine run-to-failure simulation. 2008 International Conference On Prognostics And Health Management. pp. 1-9 (2008,10), http://ieeexplore.ieee.org/document/4711414/
30. Thil, L., Read, J., Kaddah, R. & Doquet, G. Uncertainty Quantification as a Complementary Latent Health Indicator for Remaining Useful Life Prediction on Turbofan Engines. 13th IMA International Conference On Modelling In Industrial Maintenance And Reliability (MIMAR). (2025,7), https://hal.science/hal-05093810
31. Wang, S., Fan, Y., Jin, S., Takyi-Aninakwa, P. & Fernandez, C. Improved anti-noise adaptive long short-term memory neural network modeling for the robust remaining useful life prediction of lithium-ion batteries. Reliability Engineering & System Safety. 230 pp. 108920 (2023,2), https://linkinghub.elsevier.com/retrieve/pii/S095183202200535X
32. Wang, Y., Pang, Y., Chen, O., Iyer, H., Dutta, P., Menon, P. & Liu, Y. Uncertainty quantification and reduction in aircraft trajectory prediction using Bayesian-Entropy information fusion. Reliability Engineering & System Safety. 212 pp. 107650 (2021,8), https://linkinghub.elsevier.com/retrieve/pii/S0951832021001915
33. Wei, M., Ye, M., Wang, Q., Xinxin-Xu & Twajamahoro, J. Remaining useful life prediction of lithium-ion batteries based on stacked autoencoder and gaussian mixture regression. Journal Of Energy Storage. 47 pp. 103558 (2022,3), https://linkinghub.elsevier.com/retrieve/pii/S2352152X21012378
34. Zhang, C., Lim, P., Qin, A. & Tan, K. Multiobjective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics. IEEE Transactions On Neural Networks And Learning Systems. 28, 2306-2318 (2017,10), http://ieeexplore.ieee.org/document/7508982/

35. Zhao, Z., Liang, B., Wang, X. & Lu, W. Remaining useful life prediction of aircraft engine based on degradation pattern learning. Reliability Engineering & System Safety. 164 pp. 74-83 (2017,8), https://linkinghub.elsevier.com/retrieve/pii/S0951832017302454
36. Zheng, S., Ristovski, K., Farahat, A. & Gupta, C. Long Short-Term Memory Network for Remaining Useful Life estimation. 2017 IEEE International Conference On Prognostics And Health Management (ICPHM). pp. 88-95 (2017,6), http://ieeexplore.ieee.org/document/7998311/