PhD Symposium - Interpretable and Uncertainty-Aware Hybrid Prognostics Using Multimodal Knowledge for RUL Prediction

Dario Goglio[1,2,3], Dimitrios Zarouchas[1], and Manuel Arias Chao[1,2]

[1] *Delft University of Technology, 292F+VC Delft, Netherlands*
[2] *Zurich University of Applied Sciences, 8400 Winterthur, Switzerland*
[3] *Swiss International Air Lines Ltd., 8302 Kloten, Switzerland*

ABSTRACT

Unforeseen technical failures contribute significantly to airline delays, highlighting the need for predictive maintenance. However, developing reliable prognostic models in aviation is challenging due to strict safety requirements, limited labeled data, and the need for interpretable and trustworthy predictions. This research proposes a hybrid framework for remaining useful life (RUL) prediction that integrates multimodal domain knowledge available to airlines, such as sensor data, contextual information and reliability insights, into interpretable and uncertainty-aware algorithms. To this end, the proposed framework resorts to unsupervised degradation extraction with knowledge-informed autoencoders and supports extensions for failure mode segmentation. Initial experiments on a benchmark dataset show promising results, and application to real-world commercial aircraft data is planned to further validate the approach.

1. INTRODUCTION & RELATED WORK

Airline delays, which account for approximately 25% of the delays reported in 2023 (Eurocontrol, 2024), are often attributed to unforeseen technical issues, placing additional strain on already tight airline operations. In response, aviation industry has shown a growing interest in predictive maintenance. In particular, machine learning solutions for predictive maintenance have gained popularity due to the increased availability of condition monitoring data, technological advancements, and improved algorithms. It plays a crucial role in enabling predictive maintenance strategies, thus improving the efficiency of airline maintenance, leading to increased safety, reduced costs, and improved operational availability (Walthall & Rajamani, 2018; Verhagen et al., 2023).

Despite advances in purely data-driven models that directly predict RUL without providing information about the current health condition and its temporal evolution, these models are often seen as black-box models and face limited acceptance due to their lack of transparency and interpretability (Fink et al., 2020). To overcome this issue and achieve practical usability in aircraft systems, recent works such as (Guo, Li, Jia, Lei, & Lin, 2017; Kumar et al., 2022; Zhang, Zhang, Wang, Dui, & Chen, 2024) have proposed resorting to the inference of health indicators (HI) to offer a more interpretable alternative to RUL models based on direct mappings. However, developing operational prognostic models in aviation is notably challenging due to stringent safety requirements, legislative regulations, and the need for high predictive accuracy amid the scarcity of labeled data.

To mitigate the limitations posed by scarcity, Physics-Informed Machine Learning (PIML) offers a promising approach by embedding domain knowledge directly into model architectures. Karniadakis et al. (Karniadakis et al., 2021) provided a comprehensive review of PIML and categorized its strategies for embedding domain-specific knowledge into machine learning pipelines into three main types, each introducing a different kind of bias:

1. **Observational bias** integrates domain knowledge through data or feature augmentation, embedding physical behavior directly into the training data (Biggio, Bendinelli, Kulkarni, & Fink, 2023; Li et al., 2023).
2. **Inductive bias** incorporates structural or mathematical constraints into the model architecture to guide learning (Song & Liu, 2018).
3. **Learning bias** shapes the training process through customized loss functions or optimization schemes that reflect known physical principles (Qin, Yang, Zhou, Pu, & Mao, 2023).

In the context of aviation, these strategies are particularly well-suited, as airlines possess rich and diverse sources of domain knowledge. This includes time-series sensor data from onboard monitoring systems, structured reliability data such as failure rates and time-to-failure distributions, and unstructured maintenance records (commonly referred to as work orders). This work addresses the need for interpretable and uncertainty-aware prognostic models that can leverage heterogeneous sources of information available in airline operations.

Nonetheless, a critical limitation of most existing models is that they are deterministic, producing single-point predictions without accounting for uncertainty, neglecting the inherent stochasticity of the RUL estimation problem. Capturing this uncertainty is crucial for decision-making in aviation, where safety depends not only on accurate predictions but also on the quality of predictive uncertainty (Nemani et al., 2023).

2. RESEARCH GAP

In summary, this research aims to address the following gaps:

- Insufficient Integration of Reliability Data in deep learning Prognostics Model: There is a need to enhance model interpretability and predictive accuracy by incorporating reliability data.
- Lack of Interpretability in machine learning Models: Current machine learning models often lack transparency, making it difficult for domain experts to trust and act on predictions. There is a need to develop interpretable machine learning models that align with regulatory frameworks like the EASA AI Roadmap (Soudain, 2024).
- Underutilization of Contextual Information: Existing prognostic models do not fully leverage contextual information from maintenance records, work orders, and fault codes. Utilizing Large Language Models (LLMs) to extract and integrate this information can improve model performance.
- Inadequate Uncertainty Quantification: Most current deep learning methods for RUL estimation do not adequately address both epistemic and aleatoric uncertainty, which is crucial for making robust and reliable predictions in critical decision systems.

3. METHODOLOGY AND WORK PACKAGES

This research aims to address the gaps identified in section 2 by developing an uncertainty-aware hybrid framework for predicting failure times of critical sub-systems and components in commercial aircraft. The proposed framework integrates multimodal maintenance domain knowledge, including time-series sensor data, reliability information, and maintenance text records. It is hypothesized that incorporating several sources of knowledge can potentially increase failure time prediction. Additionally, combining diagnostics and prognostics into one framework, and hence increasing interpretability, will lead to higher acceptance and usability. In order to investigate the hypothesis, the following research consists of four work packages (WP).

3.1. WP1: Reliability-Informed Deep Learning Models for Uncertainty-Aware Prognostics

This work package integrates reliability theory and deep learning to develop uncertainty-aware prognostic models.

- Learn a probabilistic HI function representing component degradation, informed by reliability theory.
- Quantify uncertainty in RUL predictions to support risk-aware decision-making and trustworthiness.
- Enable extrapolation through learned HI function for interpretable RUL predictions under varying operational conditions.

3.2. WP2: Enhancing Interpretability in Prognostic Models

Improving support and acceptance from users, such as engineering departments and troubleshooting teams, is crucial for the success of prognostic models. Therefore, WP2 enhances interpretability by analyzing and researching in following areas:

- **Attention Mechanisms:** Time-feature attention models (Wang, Qin, Lu, Sun, & Shu, 2023) will be explored to highlight the most influential features and time windows contributing to degradation.
- **Failure Mode Segmentation:** Failure mode segmentation will be performed using insights from Maintenance Steering Group 3 (MSG-3) analysis, an industry-standard methodology in aviation for identifying and understanding failure modes. By clearly representing the degradation state of each failure mode, this approach supports more precise maintenance actions and long-term repair strategies.
- **Interpretability Tooling and LLMs:** Tools such as Pyreal (Zytek, Wang, Liu, Berti-Equille, & Veeramachaneni, 2023) will be explored to explain model predictions, and LLMs will be used to generate user-friendly output explanations for maintenance engineers (Zytek, Pidò, & Veeramachaneni, 2024).

3.3. WP3: Leveraging Contextual Information for Prognostics with LLM

The objective of this work package is to improve prognostic models by leveraging contextual information through LLMs.

- LLMs will be trained to extract relevant insights from structured and unstructured sources (e.g., fault codes, maintenance manuals, work orders).
- Extracting and classifying failure modes based on maintenance records
- Integrating contextual insights such as health index adjustments following maintenance actions

3.4. WP4: Transfer Learning for Cross-fleet Prognostics with Multimodal Domain Knowledge-based Prognostics Models

This WP focuses on adapting the developed models to other fleets with limited data.

- Transfer learning will be applied to adapt models developed on a data-rich source fleet (e.g., Airbus A220) to a data-sparse target fleet (e.g., Airbus A320neo).
- Techniques such as fine-tuning and domain adaptation will be employed. Data augmentation may be used to compensate for sparsity in the target domain.

4. RESULTS & ANALYSIS

In this study, HIs are considered a stochastic definition of health over time, incorporating reliability information via Weibull distribution parameters derived from fleet reliability information. The HI is expressed as (Dersin, Bajarunas, & Chao, 2024):

$$h(t) = 1 - bt^p \tag{1}$$

where p is a fleet-wide shape parameter and b is a unit-specific random variable modeled via a 2-parameter Fréchet (λ_b,β_b) distribution, derived from Weibull (β, η) time-to-failure priors.

4.1. Reliability-Informed Deep Learning (RIDL)

Three variants of the RIDL framework are developed to embed reliability knowledge into deep learning models:

- **RIDL-VAE:** A variational autoencoder (VAE) embedding $b \sim$ Fréchet(λ_b, β_b) in the latent space, regularized via a closed-form KL divergence derived for the Fréchet distribution. The sampled b is then used to compute the analytical HI curve (see eq. 1), which serves as input to the decoder, hence introducing reliability-informed inductive bias. Note that the exponent p was assumed to be known.
- **RIDL-AAE:** An adversarial autoencoder (AAE) (Bermejo-Barbanoj, Moya, Badías, Chinesta, & Cueto, 2024) embedding the random variable b in the latent space. A discriminator enforces alignment with the Fréchet prior, introducing inductive bias without explicit likelihood assumptions, similar in architecture to RIDL-VAE. A secondary AAE variant was also explored, embedding the health index directly in the latent space. However, lacking an analytical degradation curve, this approach does not support extrapolation and was therefore not evaluated.
- **RIDL-AE:** A conditional autoencoder, inspired by (Bajarunas, Baptista, Goebel, & Chao, 2023), where the health index is embedded in the latent space. Initial RIDL models assumed a constant, known exponent p. However, this is unrealistic as degradation curves are normally not known beforehand. To address this, the current model extends the RIDL framework by learning both the health index and the degradation shape function $p = p(s)$ in parallel, where p exponent depends on the degradation level s. The health index is then given by:

$$h(t) = 1 - \left(\frac{t}{\eta(-\log q)^{\frac{1}{\beta}}} \right)^{p(s)} \tag{2}$$

This formulation is equivalent to the analytical model (Eq. 1), where the random variable b is derived from reliability theory (Dersin et al., 2024). This hybrid model jointly learns unit-specific degradation parameters q and the fleet-wide shape function $p(s)$, allowing both fleet-level generalization and unit-level adaptation. It leverages reliability-based priors to regularize health index estimation during training, while supporting extrapolation during inference.

4.2. Experimental Setup

Experiments were performed on the N-CMAPSS DS03 dataset (Arias Chao, Kulkarni, Goebel, & Fink, 2021), which simulates turbofan degradation under realistic flight conditions and closely mirrors Quick Access Recorder (QAR) sensor data used by airlines. Following frameworks were compared:

- **RIDL-AE**: with learned $p(s)$ and extrapolated HI.
- **RIDL-AAE**: with fixed p and statistical prior on b.
- **Supervised Model**: trained on sensor-to-RUL mappings (Arias Chao, Kulkarni, Goebel, & Fink, 2022)

Prediction is performed at fixed observation cycles, and the MAE of the RUL estimates is reported.

Cycle	RIDL-AE	RIDL-AAE	Supervised Model
5	41.22	6.83	7.66
15	30.67	6.17	7.08
30	9.08	6.33	8.66
45	**5.25**	6.50	6.68
Average	21.55	**6.46**	7.53

Table 1. Mean Absolute Error (MAE) for RUL predictions made at different observation cycles.

4.3. Discussion

RIDL-AE achieves the lowest MAE at cycle 45, indicating strong performance as the system nears failure. However, its early-cycle errors are higher, partly attributed due to numerical instability in estimating q_s via:

$$q_s = e^{-\left(\frac{t_s}{(1-s)^{\frac{1}{p}}\eta} \right)^{\beta}} \tag{3}$$

At early stages (t_s small, s high), the denominator $(1-s)^{1/p}$ becomes small, amplifying the exponent and resulting in extremely small q_s values. This leads to overestimated time-to-failure and thus poor RUL predictions. Incorporating uncertainty quantification is expected to mitigate this effect. In fact, even median-based ($q_s = 0.5$) predictions achieve MAE around 9.

RIDL-AAE, with a fixed p, is more stable in early predictions but less adaptive closer to failure. This contrast suggests that RIDL-AE's latent space representation of the health index allows more nuanced extraction of degradation patterns, compared to the fixed parameterization in RIDL-AAE, offering a promising direction for further improvement.

5. CONCLUSION

Integrating reliability theory into deep learning, this research improves the transparency and accuracy of predictive maintenance RUL predictions. By embedding domain-informed priors and modeling the health index analytically, it addresses key challenges such as sparse failure data and varying operational conditions.

Notably, RIDL-AE demonstrates strong performance at later degradation stages, while RIDL-AAE provides stable early-stage predictions through structured priors. This highlights the value of combining inductive and learning biases.

Ongoing work aims on enhancing early-cycle predictions through uncertainty modeling and extending the framework to real-world datasets from commercial aircraft. The modularity of the RIDL architecture supports future extensions for failure mode classification and deployment in operational decision-support tools, advancing reliable, interpretable prognostics for safety-critical domains.

REFERENCES

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021, January). Aircraft Engine Run-to-Failure Dataset under Real Flight Conditions for Prognostics and Diagnostics. *Data*, *6*(1), 5.

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2022, January). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Bajarunas, K., Baptista, M., Goebel, K., & Chao, M. A. (2023, October). Generic Hybrid Models for Prognostics of Complex Systems. *Annual Conference of the PHM Society*, *15*(1).

Bermejo-Barbanoj, C., Moya, B., Badías, A., Chinesta, F., & Cueto, E. (2024, February). *Thermodynamics-informed super-resolution of scarce temporal dynamics data.*

Biggio, L., Bendinelli, T., Kulkarni, C., & Fink, O. (2023, September). Ageing-aware battery discharge prediction with deep learning. *Applied Energy*, *346*, 121229.

Dersin, P., Bajarunas, K., & Chao, M. A. (2024). Analytical Modeling of Health Indices for Prognostics and Health management.

Eurocontrol. (2024, March). *Annual Network Operations Report 2023 | EUROCONTROL* (Tech. Rep.).

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020, June). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017, May). A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, *240*, 98–109.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021, June). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Kumar, A., Parkash, C., Vashishtha, G., Tang, H., Kundu, P., & Xiang, J. (2022, May). State-space modeling and novel entropy-based health indicator for dynamic degradation monitoring of rolling element bearing. *Reliability Engineering & System Safety*, *221*, 108356.

Li, X., Teng, W., Peng, D., Ma, T., Wu, X., & Liu, Y. (2023, May). Feature fusion model based health indicator construction and self-constraint state-space estimator for remaining useful life prediction of bearings in wind turbines. *Reliability Engineering & System Safety*, *233*, 109124.

Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., ... Hu, C. (2023, September). *Uncertainty Quantification in Machine Learning for Engineering Design and Health Prognostics: A Tutorial.* arXiv.

Qin, Y., Yang, J., Zhou, J., Pu, H., & Mao, Y. (2023, April). A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery RUL prediction. *Advanced Engineering Informatics*, *56*, 101973.

Song, C., & Liu, K. (2018, October). Statistical degradation modeling and prognostics of multiple sensor signals via data fusion: A composite health index approach. *IISE Transactions*, *50*(10), 853–867.

Soudain, G. (2024, March). *EASA Concept Paper: Guidance for Level 1 & 2 machine learning applications.*

Verhagen, W. J. C., Santos, B. F., Freeman, F., van Kessel, P., Zarouchas, D., Loutas, T., ... Heiets, I. (2023, September). Condition-Based Maintenance in Aviation: Challenges and Opportunities. *Aerospace*, *10*(9), 762.

Walthall, R., & Rajamani, R. (2018, September). The Role of PHM at Commercial Airlines. In M. G. Pecht & M. Kang (Eds.), *Prognostics and Health Management of Electronics* (1st ed., pp. 503–534). Wiley.

Wang, Q., Qin, K., Lu, B., Sun, H., & Shu, P. (2023, August). Time-feature attention-based convolutional autoencoder for flight feature extraction. *Scientific Reports*, *13*(1), 14175.

Zhang, Y., Zhang, C., Wang, S., Dui, H., & Chen, R. (2024, January). Health indicators for remaining useful life prediction of complex systems based on long short-term memory network and improved particle filter. *Reliability Engineering & System Safety*, *241*, 109666.

Zytek, A., Pidò, S., & Veeramachaneni, K. (2024, May). *LLMs for XAI: Future Directions for Explaining Explanations.*

Zytek, A., Wang, W.-E., Liu, D., Berti-Equille, L., & Veeramachaneni, K. (2023, December). *Pyreal: A Framework for Interpretable ML Explanations.* arXiv.

BIOGRAPHIES

Dario Goglio is a PhD candidate at Delft University of Technology (Netherlands) and Zurich University of Applied Sciences (Switzerland). He holds an MSc in Mechanical Engineering from ETH Zurich and is working for Swiss International Air Lines as an Aircraft Data Analytics Engineer, where he gains valuable domain knowledge and the opportunity to apply his research in real-world scenarios. His research focuses on developing hybrid models for prognostics using multimodal domain knowledge available to airlines.

Dimitrios Zarouchas is an Associate Professor in the Aerospace Structures & Materials Department at Delft University of Technology in the Netherlands. He leads the PYTHIA research team on Artificial Intelligence for Structures, Prognostics, and Health Management, and is the founder of the Center of Excellence in AI for Structures. His work focuses on developing intelligent systems for real-time diagnostics and prognostics of lightweight structures used in aviation, space, wind energy, and naval industries.

Manuel Arias Chao is a Senior Lecturer at Zurich University of Applied Sciences and an Assistant Professor at the Operations and Environment Chair of the Delft University of Technology in the Netherlands. He has a PhD in Physics-informed Machine Learning for Prognostics and Health Management from ETH Zurich, a Master's degree in Thermal Power from Cranfield University, and a Bachelor's degree in Aeronautical Engineering from the Technical University of Madrid. Manuel has gained valuable industrial and research experience as a visiting researcher at the Diagnostics & Prognostics Group at NASA Ames, Thermodynamics & Performance Lead Engineer at General Electric and ALSTOM Power, and Aero Engine Maintenance Engineer at ITP. In his current role, Manuel also co-leads the Expert Group Smart Maintenance from the Swiss Alliance for Data-Intensive Services.