Rethinking RUL Prediction: Uncertainty, Robustness, Interpretability, and Feasibility Matter

Mariana Salinas-Camus[1], Kai Goebel[2], and Nick Eleftheroglou[3]

[1, 3] *Intelligent System Prognostics Group, Aerospace Structures and Materials Department, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, Delft, 2629HS, the Netherlands*
m.salinascamus@tudelft.nl
n.eleftheroglou@tudelft.nl

[2] *Fragum Global, Mountain View, CA 94040, USA*
Lulea University of Technology, Lulea, 971 87, Sweden
kai.goebel@rocketmail.com

ABSTRACT

Prognostics and Health Management (PHM) plays a key role in predicting the Remaining Useful Life (RUL) of systems, which is essential for enabling decision-making for Predictive Maintenance (PdM) and operations. While most research has traditionally focused on improving the accuracy of RUL predictions, this paper argues that four essential characteristics, uncertainty, robustness, interpretability, and feasibility, are key for real-world PHM applications. This study explores these characteristics through a comparative analysis of two data-driven models (DDMs): the probabilistic Bidirectional Long Short-Term Memory (BiLSTM) model and the Adaptive Hidden Semi-Markov Model (AHSMM). Deep Learning (DL) models such as the BiLSTM often achieve high prediction accuracy but struggle with uncertainty quantification and adaptability across varying operating conditions. In contrast, stochastic models like AHSMM offer stronger robustness and feasibility, performing well even with limited or noisy data. Using the C-MAPSS dataset, the models are evaluated through the lens of the four proposed characteristics. This more holistic approach clarifies each model's strengths, limitations, and practical trade-offs in PHM settings. The findings highlight that while accuracy remains important, focusing solely on it can overlook critical factors that affect model performance in real operational environments. Balancing all four characteristics is essential for deploying reliable and effective decision-making for predictive maintenance and operations.

1. INTRODUCTION

Prognostics and Health Management (PHM) analyzes current and future health conditions of engineering systems to improve reliability, reduce maintenance costs, and ensure safety through interconnected processes of feature extraction, diagnostics, prognostics, and decision-making. Prognostics, the most challenging PHM aspect, predicts Remaining Useful Life (RUL) to enable Predictive Maintenance (PdM) (Z. Huang et al., 2017), shifting maintenance from reactive or time-based approaches to PdM planning that anticipates failures and optimizes repair scheduling.

However, effective prognostic integration into real-world applications requires more than accurate predictions (Zio, 2022). Prognostic models must address four essential characteristics: uncertainty quantification (UQ) to help decision-makers assess risk despite inherent degradation unpredictability (Kamariotis et al., 2024), robustness to ensure reliable performance across varying operational scenarios (C. Huang et al., 2024), interpretability to provide decision-making transparency required by regulations (Goodman & Flaxman, 2017), and feasibility to function effectively with realistic data constraints (Kamariotis et al., 2024). These characteristics collectively enable trustworthy, resilient, and actionable predictions essential for successful PdM frameworks.

This work examines the importance of these four characteristics in developing data-driven prognostic models. While traditional evaluation criteria like prediction accuracy remain important, they are insufficient on their own. Real-world deployment demands models that not only predict well but also do so in a way that supports complex operational decisions. For instance, a highly accurate model that cannot express prediction confidence or requires impractical amounts of labeled training data may not be viable in practice.

To explore these concepts, this paper investigates data-driven prognostic models, particularly Deep Learning (DL) and Stochastic Models, building on the models reviewed in the earlier comprehensive journal publication (Salinas-Camus et al., 2025). The objective is to understand how different modeling paradigms address the key requirements for effective decision-making in PdM.

2. KEY CHARACTERISTICS IN PROGNOSTICS

Effective prognostic models for predictive maintenance require a careful balance of several essential characteristics beyond just high predictive accuracy. These models must also quantify inherent uncertainties in RUL predictions, maintain robustness across diverse operational conditions, offer interpretability to meet regulatory and practical demands, and demonstrate feasibility for deployment in real-world industrial systems. Together, these characteristics ensure that prognostic models deliver both technical excellence and practical utility within comprehensive PdM frameworks.

2.1. Uncertainty

Uncertainty Quantification (UQ) in prognostics assesses how various sources of uncertainty affect RUL predictions, which are inherently uncertain and modeled as random variables. Uncertainties are categorized as aleatoric (data randomness) and epistemic (lack of knowledge) (Der Kiureghian & Ditlevsen, 2009), with this information being crucial for informed decision-making in health management systems.

UQ remains a significant challenge for DL-based prognostic models due to their deterministic nature (Vollert & Theissler, 2021), which typically produces point estimates rather than uncertainty-aware predictions. Various Bayesian techniques have been developed to address this, including Variational Inference, Monte Carlo (MC) Dropout, Deep Gaussian Processes, and Markov Chain Monte Carlo methods, each with trade-offs in accuracy and computational complexity (Abdar et al., 2021). MC Dropout (Gal & Ghahramani, 2016) is widely adopted for its simplicity but suffers from poor approximation of complex posteriors (Fort et al., 2019) and sensitivity to user-defined parameters (Folgoc et al., 2021; Salinas-Camus & Eleftheroglou, 2024). Most Bayesian DL implementations in prognostics still fall short in fully capturing uncertainty, often using suboptimal dropout rates that result in underestimated uncertainty and narrow confidence intervals (Pei et al., 2022; Zhu et al., 2022). While advanced approaches like concrete dropout (Lin & Li, 2022) and ensemble methods (Alcibar et al., 2024) attempt to improve calibration, they still suffer from assumptions like Gaussian distributions that may not reflect real-world prognostic complexities.

Stochastic models naturally account for aleatoric uncertainty by modeling inherent randomness in degradation processes using probabilistic approaches (Xie et al., 2016), typically providing closed-form posterior distributions that allow confidence intervals to be derived directly. However, these models often produce wide confidence intervals, given that it is the aleatoric uncertainty that arises from the phenomenon, as seen in Non Homogeneous Hidden Semi Markov Models (NHHSMMs) for turbofan engines (Moghaddass & Zuo, 2014) and composite specimens (Eleftheroglou & Loutas, 2016). Enhanced approaches like Similarity Learning Hidden Semi-Markov Model (SLHSMM) (Eleftheroglou et al., 2024) and unit-to-unit (N. Li et al., 2022) adaptive frameworks address this by incorporating data similarity or selecting optimal models for individual cases, effectively managing uncertainty to improve prediction reliability. While stochastic models excel at modeling aleatoric uncertainty, they generally do not capture epistemic uncertainty unless specifically extended, such as the Generalized Hidden Markov Model (GHMM) (Xie et al., 2016), which quantifies both uncertainty types using imprecise probabilities, though at high computational cost.

2.2. Robustness

Robustness in prognostics models refers to maintaining reliable predictions despite variations in operational conditions, environmental factors, and input data quality. The primary challenge occurs when testing conditions differ from training data, creating domain shifts that degrade performance. While Domain Adaptation techniques have been developed to address these shifts (da Costa et al., 2020; X. Li et al., 2020; Vollert & Theissler, 2021), guaranteeing robustness remains a significant challenge for system safety and reliability.

Several DL approaches have been proposed with limited success. Deep Convolutional Neural Networks (CNNs) with adaptive batch normalization for turbofan engines (J. Li et al., 2019) and domain adversarial neural networks for turbofan engines and bearings (da Costa et al., 2020; X. Li et al., 2020) showed improvements but still exhibited high errors and volatility, with performance constrained by data scarcity in industrial scenarios. Transfer Learning frameworks (Zhang et al., 2021) face limitations from assuming straight degradation patterns and requiring "healthy" state data, which is often unavailable in industrial applications.

Stochastic models, particularly Hidden Markov Models (HMMs), show promise through adaptive methodologies. The Adaptive Non-Homogeneous Hidden Semi Markov Model (ANHHSMM) successfully adapts to loading changes such as impact loading without requiring large datasets, outperforming non-adaptive models on open-hole specimens (Eleftheroglou et al., 2020). However, it has only been validated for brief operational changes, not sustained condition changes. The SLHSMM addresses similar challenges by characterizing similarity between testing and training

data, demonstrating superior performance against outliers while reducing confidence intervals and computational costs (Eleftheroglou et al., 2024). Nevertheless, its effectiveness depends heavily on diverse outliers in the training set, limiting its applicability in data-scarce environments.

2.3. Interpretability

Interpretability is used in this work as an umbrella term that encompasses both the understanding of the internal mechanisms of prognostic models (often referred to as interpretability in a narrow sense) and the generation of human-understandable explanations for model outputs (commonly referred to as explainability) (Kobayashi & Alam, 2024). This aspect has gained importance due to the increased use of "black box" DL models and regulatory requirements, such as the General Data Protection Regulation (GDPR) (Goodman & Flaxman, 2017), which grants the right to explanation for algorithmic decisions affecting individuals. In PHM applications involving safety-critical components, interpretability in this broader sense is considered essential for informed decision-making and regulatory compliance (Sharma et al., 2024).

Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are widely used interpretability tools for DL-based prognostic models, focusing on feature selection and importance analysis (Figueroa Barraza et al., 2021; Baptista et al., 2024; Serradilla et al., 2020). However, both exhibit significant limitations in robustness and consistency. LIME produces unreliable explanations that can vary drastically for identical inputs (Alvarez-Melis & Jaakkola, 2018; Garreau, 2023), while SHAP relies on approximations that can diverge substantially from exact values and may assign misleading feature importance (Ali et al., 2023). These tools often produce conflicting feature rankings, creating a "disagreement problem" that complicates decision-making in critical PHM applications. While approaches as trust scores (Kundu & Hoque, 2023) have been proposed to identify suitable explanation tools, they remain limited and do not address the root causes of explanation biases. Other DL interpretability approaches include variational encoders creating interpretable latent spaces (Costa & Sánchez, 2022) and Relevant Vector Machines (Alamaniotis, 2023), but these methods provide incomplete frameworks for model interpretation since they focus on latent-space visualization or general signal characteristics, without explaining how specific inputs influence model predictions. Despite extensive research efforts, DL models remain fundamentally black boxes, with interpretability tools still facing significant challenges in providing reliable explanations.

Stochastic models, in contrast, offer inherent interpretability advantages through their probabilistic structure without requiring additional explanation tools. HMMs can directly represent physical degradation stages through hidden states that correspond to actual damage progression (matrix cracking, delamination, fiber breakage, failure) (Loutas et al., 2017), predict both RUL and current damage states simultaneously, and provide intuitive insights into degradation processes through interpretable parameters such as state transition probabilities and average time spent in each damage state. Enhanced variants, namely, the SLHSMM (Eleftheroglou et al., 2024) and the ANHHSMM (Eleftheroglou et al., 2020) further improve interpretability through similarity estimation and outlier identification, offering users a deeper understanding of both predictions and underlying data compared to black box DL approaches.

2.4. Feasibility

Feasibility refers to a model's ability to be trained and achieve reliable results with the available data, which varies significantly by industry and often involves limitations in data quantity, quality (Zio, 2022; Gebraeel et al., 2023), diversity (Verhagen et al., 2023), or labeling. This challenge is particularly pronounced in industrial applications where labeled data is scarce or missing measurements make it difficult to create comprehensive training datasets for prognostic models.

DL models face significant feasibility challenges due to their requirement for large, labeled datasets (Vollert & Theissler, 2021). To address limited data scenarios, "few-shot prognostics" approaches have emerged, including Bayesian approximation enhanced probabilistic meta-learning (BA-PDL) that provides interval estimates and uncertainty quantification for variable-length predictions, though results can be noisy with higher RMSE as sample size increases (Ding et al., 2023). Graph neural networks create dynamic graphs to uncover hidden patterns and predict remaining lifespan even under changing conditions (Ding et al., 2024), while Bayesian DL approaches tackle unlabeled data by preprocessing monitoring data to create degradation-labeled samples and employing variational inference for UQ (Pei et al., 2022). Self-supervised LSTM-CNN frameworks use contrastive learning to extract features from raw sensor signals, reducing dependence on labeled data, though they incur high computational overhead that limits deployment in resource-constrained settings (Deng et al., 2024).

In contrast, stochastic models demonstrate superior feasibility as they can be effectively trained with small datasets without requiring labeled data. The NHHSMM has been successfully trained using only 8 degradation histories (Eleftheroglou & Loutas, 2016), while advanced variants such as ANHHSMM and SLHSMM are particularly well-suited for few-shot prognostics, capable of adapting to or utilizing just a single degradation history. Additionally, non-parametric stochastic models based on functional principal component analysis can handle fragmented data with miss-

ing sensor readings, maintaining reliable RUL predictions even with up to 25% of data missing, demonstrating their robustness in data-scarce industrial environments (N. Li et al., 2024).

3. CASE STUDY

This case study compares two prognostic models, one DL-based and one stochastic, across four key characteristics. The DL model is a Bidirectional Long Short-Term Memory (BiLSTM) network, known for its ability to capture temporal dependencies. Inspired on the approach in (Caceres et al., 2021), it uses a probabilistic framework for predicting the RUL while capturing the aleatoric and epistemic uncertainty.

The stochastic model is the Adaptive Hidden Semi-Markov Model (AHSMM), chosen for its robustness in generalizing to unseen degradation patterns, following the framework in (Eleftheroglou et al., 2020).

Both models are evaluated using the C-MAPSS FD001 sub-dataset, which includes one fault mode and one operational condition. To ensure full RUL labels are available, only complete run-to-failure trajectories are used, 64 for training and 16 for testing, enabling consistent evaluation of metrics like uncertainty coverage across the entire life cycle. Hyperparameters for both models are optimized on this baseline data.

Model performance is assessed using Root Mean Square Error (RMSE), standard deviation of RMSE across test instances, uncertainty coverage, and Continuous Ranked Probability Score (CRPS) (Hersbach, 2000), with different experiments conducted to evaluate all four key characteristics.

3.1. Prognostics Models

The following subsections briefly describe the mathematical modeling and architecture of each prognostic model used in this study.

3.1.1. Adaptive Hidden Semi-Markov Model

The Hidden Semi-Markov Model (HSMM) extends the classic HMM (Rabiner, 1989), by explicitly modeling the sojourn time, i.e., how long the system remains in each damage state, using a Weibull distribution. Each state S_i emits observations for a duration sampled from the Weibull distribution.

The model assumes left-to-right transitions (no recovery), starts from a healthy state, and ends in a known failure state. Parameters are learned via the Expectation-Maximization (EM) algorithm, and each damage state corresponds to a level of degradation. Due to the model's constraint of being trained on a single feature, sensor s_{11} is selected as input based on its superior prognosability, trendability, and monotonicity (Coble & Hines, 2009). The optimal number of states (seven for this case study) is determined using the Bayesian Information Criterion (BIC) (Moghaddass & Zuo, 2014).

The AHSMM enhances the HSMM by adjusting future state durations based on observed transitions. When a transition from state S_i to S_{i+1} occurs, the observed sojourn time T_i is compared to its expected value E_i, and their ratio defines the resampling factor R_f. This factor updates the scale parameter for the next state's Weibull distribution, as shown in Equation 1. By adapting the scale parameter (while keeping the shape parameter fixed), the Weibull distribution better represents the sojourn time when the degradation process has changed.

$$\beta^*_{i+1} = \frac{E_{i+1} R_f}{\Gamma\left(1 + \frac{1}{\alpha_{i+1}}\right)} \tag{1}$$

To estimate RUL, a probabilistic time-dependent measure is used (Kontogiannis et al., 2025), accounting for time τ already spent in the current state. The probabilities of staying in or transitioning out of the current state are:

$$d_{i,i+1} = P(d \leq \tau \mid S_t = i), \quad d_{i,i} = 1 - d_{i,i+1}. \tag{2}$$

The final RUL distribution, which includes future state durations and a Gaussian noise term $\mathcal{N}(1, \epsilon)$, is given by:

$$\begin{aligned} Pr(RUL_i^t) = d_{i,i} \left(D_i(d - \tau) + \sum_{k=i+1}^{N-1} D_k(d) + \mathcal{N}(1, \epsilon) \right) \\ + d_{i,i+1} \left(\sum_{k=i+1}^{N-1} D_k(d) + \mathcal{N}(1, \epsilon) \right). \end{aligned} \tag{3}$$

This formulation yields a full probability distribution over RUL, not just a point estimate. The 95% confidence intervals are computed from the cumulative distribution function, enabling prognostics with aleatoric UQ.

3.1.2. Probabilistic Bidirectional Long Short-Term Memory network

A model based on a Bidirectional Long Short-Term Memory (BiLSTM) network was developed, inspired by the work of (Caceres et al., 2021), as mentioned before. The input data consisted of multivariate time series from selected engine sensors that show high correlation values (s_2, s_4, s_7, s_{11}, s_{12}, s_{15}, s_{17}, s_{20}, and s_{21}), along with the corresponding RUL values (for an analysis with only one sensor see (Salinas-Camus et al., 2025)). The data was segmented into overlapping windows of 10 timesteps with a stride of 1, allowing for online use. Each window was labeled using the

RUL value at its final timestep. If a window contained the end of an engine's life (RUL = 0), it was padded using the last valid row to maintain uniform shape across samples. Sensor features and RUL values were normalized using separate standard scalers.

The model architecture includes two stacked Bidirectional LSTM layers, each with 100 units, followed by dropout layers with a rate of 0.2. These values are derived from a Bayesian optimization performed with Keras Tuner. These dropout layers served both as regularization during training and as a mechanism for uncertainty estimation via MC dropout at inference time. The model produces two outputs: the predicted mean RUL and a predicted standard deviation, which, after being passed through a softplus activation, represents the aleatoric uncertainty. To additionally estimate epistemic uncertainty, 100 stochastic forward passes are performed during inference using MC dropout.

Training was performed using the Adam optimizer with a learning rate of 1e-4. The loss function, shown in Equation 4, is the negative log-likelihood of the Gaussian predictive distribution with a regularizer to encourage well-calibrated predictions. Early stopping and learning rate reduction on a plateau were used to prevent overfitting and improve convergence.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\mathcal{N}(y_i|\mu_i,\sigma_i^2) + \lambda\cdot\frac{1}{N}\sum_{i=1}^{N}(\log(\sigma_i))^2 \quad (4)$$

To estimate epistemic uncertainty, 100 stochastic forward passes are performed during inference using MC dropout. The mean and standard deviation of the predictions were used to compute 95% confidence intervals. The final evaluation included denormalizing the predicted RUL values.

3.2. Results

This section presents the results for both models using the baseline case, followed by an analysis of the four key characteristics using targeted experiments to evaluate model behavior.

3.2.1. Baseline

The baseline case uses subdataset FD001, featuring a single operational condition and fault mode. Figure 1 shows overlapping lifetime distributions for training and test sets, indicating well-aligned distributions.

Table 1 summarizes baseline results: AHSMM achieves an RMSE of 34.48, while BiLSTM slightly outperforms it with 33.65. However, BiLSTM predictions have higher variability, reflecting less consistency.

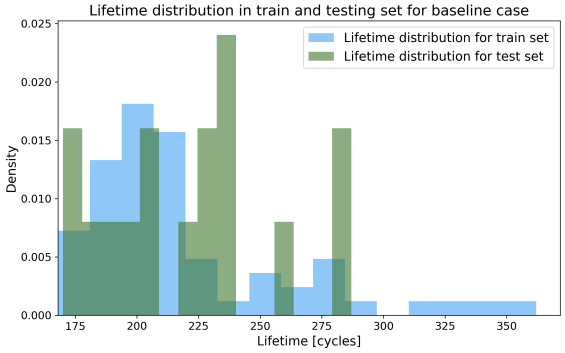

Figure 1. Lifetime distribution for baseline case.

Figure 2 illustrates predictions for engine 7. BiLSTM shows greater volatility, likely due to sensitivity to input fluctuations, whereas AHSMM provides smoother but occasionally jumpy predictions from discrete state transitions. Both models improve accuracy in the final cycles as degradation becomes clearer.

Model	RMSE	SD
BiLSTM	33.65	19.29
AHSMM	34.48	9.34

Table 1. Results of prognostic models for baseline case.

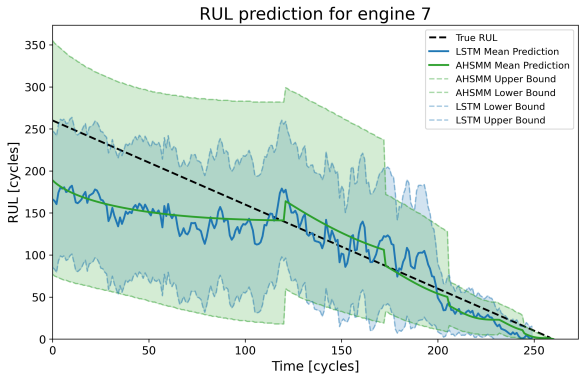

Figure 2. RUL predictions for engine 7 in baseline case.

3.2.2. Uncertainty

UQ remains a major challenge in prognostics, particularly when comparing models that account for different sources of uncertainty. In this section, the coverage metric, which measures how often true RUL values fall within predicted intervals, and CRPS, which assesses the accuracy and sharpness of the full predictive distribution, are used to evaluate uncertainty.

The coverage metric is defined as the proportion of true values that fall within the predicted confidence bounds. For each

time step t, if y_t is the true RUL and $[l_t, u_t]$ is the predicted 95% confidence interval, coverage is calculated as:

$$C_t = \begin{cases} 1, & \text{if } l_t \leq y_t \leq u_t \\ 0, & \text{otherwise} \end{cases}, \quad \text{Coverage} = \frac{1}{T}\sum_{t=1}^{T} C_t \quad (5)$$

As shown in Table 2, both models achieve high coverage, indicating reliable intervals. However, BiLSTM yields a lower CRPS, suggesting sharper and more accurate distributions. This reflects better-calibrated uncertainty estimates, partly due to its modeling of both aleatoric (via predicted mean and SD) and epistemic uncertainty (via MC dropout). In contrast, AHSMM captures only aleatoric uncertainty.

Despite high reliability, the wide intervals highlight the need to consider interval width alongside coverage and CRPS to ensure predictions remain actionable.

Model	Cov.	CRPS
BiLSTM	0.98	14.76
AHSMM	0.97	19.49

Table 2. Coverage results of the prognostics models for the baseline case.

3.2.3. Robustness

Robustness was assessed by evaluating the models' generalization ability when tested on unseen fault modes. Both models were trained on the FD001 dataset, which includes only HPC degradation, and tested on the FD003 dataset, which introduces an additional fault mode (fan degradation) under the same operating conditions. This mismatch in degradation mechanisms introduces a distribution shift, as shown in Figure 3 that challenges the models' adaptability.

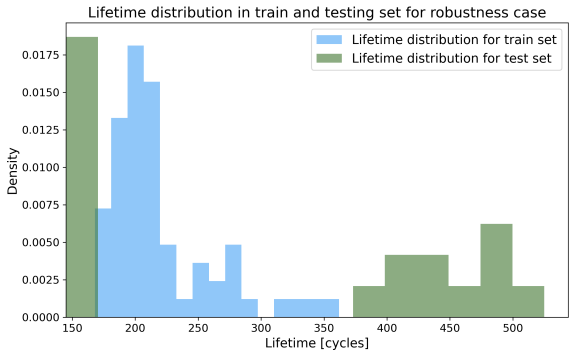

Figure 3. Lifetime distribution for robustness case.

Table 3 shows that both models experience a significant increase in RMSE when exposed to this new fault mode, indicating degraded predictive accuracy. BiLSTM sees a 146.39% increase in RMSE, while the AHSMM shows a slightly lower decline in performance at 137.47%.

Model	RMSE	ΔRMSE (%)	SD
BiLSTM	82.91	+146.39%	49.67
AHSMM	81.88	+137.47%	51.60

Table 3. Results of prognostic models for robustness case.

Model	Cov. (%)	ΔCov. (%)	CRPS	ΔCRPS (%)
BiLSTM	0.72	-26.53%	14.72	+0.27%
AHSMM	0.82	-15.46%	46.26	+137.35%

Table 4. Results of prognostics models in terms of uncertainty for robustness case.

Despite the drop in accuracy, the coverage metric in Table 4 indicates that both models still provide reasonably reliable uncertainty estimates. AHSMM achieves a higher coverage (0.82 vs. 0.72 for BiLSTM), although both show reduced performance compared to their baselines. This is expected, as the models are exposed to a fault type not seen during training.

Regarding CRPS, the BiLSTM maintains a low and stable score (+0.27% change), suggesting its probabilistic predictions remain relatively well-calibrated and sharp despite the challenging conditions. In contrast, the AHSMM's CRPS deteriorates substantially (+137.35%), indicating less accurate UQ under the new fault mode.

Figures 4 (shorter lifetime with respect to training) and 5 (longer lifetime with respect to training) visualize the RUL predictions of both models. Notably, in Figure 5, the BiLSTM model fails to converge to a RUL of zero. As a supervised model, the BiLSTM struggles with data that diverges from the training distribution, leading to unreliable predictions toward the end of the lifetime—particularly for engines with longer operational durations than those seen during training. This drop in accuracy near the end of life is particularly concerning for critical applications, where precise end-of-life prediction is essential.

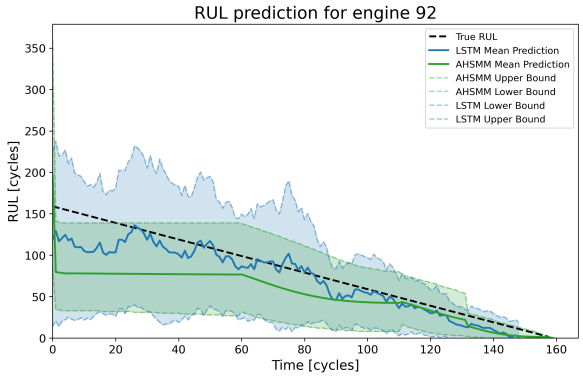

Figure 4. RUL predictions for engine 92.

On the other hand, AHSMM demonstrates slightly better robustness in terms of both RMSE and UQ, with more consistent predictions through the lifetime of the engine and higher

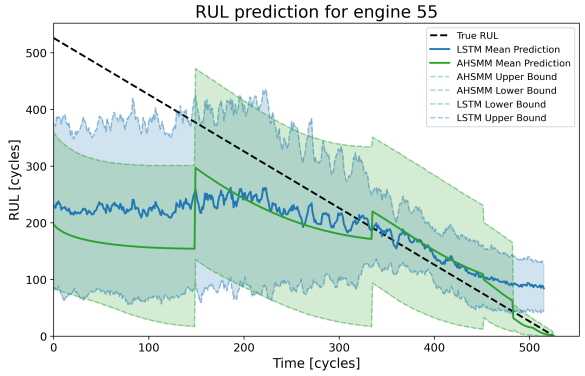

Figure 5. RUL predictions for engine 55.

coverage. Its performance under this distribution shift suggests it may be more suitable for applications that might contain unseen data during online operation.

3.2.4. Interpretability

The AHSMM offers a relatively interpretable approach to RUL prediction, thanks to its structured design and reliance on the Weibull distribution, which is commonly used in reliability analysis. The scale parameter of the Weibull distribution is dynamically adjusted based on observed transitions, allowing the model to adapt to real-time degradation behavior. This adaptive mechanism creates a direct and intuitive link between predictions and observed data.

Figure 6 shows the sojourn time Weibull distributions for hidden states 5 and 6 for engine 37, which has a short lifetime of 170 cycles, making it a left outlier in the overall distribution (see Figure 1). The adaptive mechanism, depicted by the dashed lines, shifts the distributions to the left to account for the shorter time spent in each state.

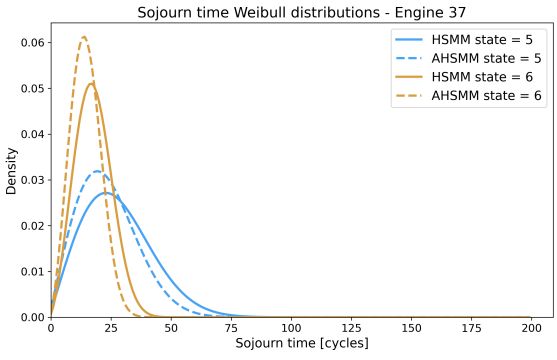

Figure 6. Sojourn time Weibull distributions for hidden states 5 and 6 for engine 37 RUL prediction.

In addition, the AHSMM provides visual representations of degradation progression, enhancing transparency. Each hidden state corresponds to a degradation level, and transitions between states can be tracked and interpreted. Figure 7 shows the sequence of estimated states for engine 37, which can be used as a diagnostic tool to assess the system's condition.

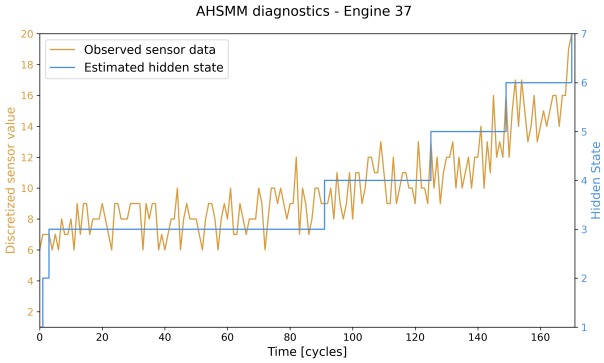

Figure 7. Diagnostics of engine 37's degradation levels based on hidden state estimates.

In contrast, the BiLSTM model, with 32,9202 trainable parameters, operates as a black box. While powerful, its predictions are less interpretable and difficult to trace back to specific features or time points. However, interpretability can be partially recovered through post hoc analysis. Figure 8 presents a SHAP-based heatmap of feature importance over time, averaged across all samples. Sensors 4, 11, and 12 emerge as the most influential, with sensor 11 peaking near the end of life. Interestingly, the same sensor is used exclusively by the AHSMM due to its high values for prognostics metrics (i.e., monotonicity, trendability, and prognosability).

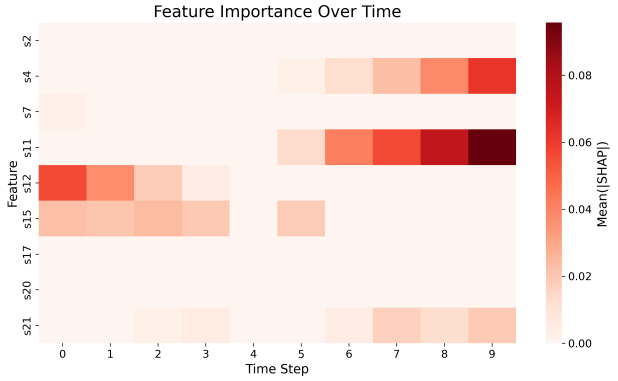

Figure 8. SHAP-based heatmap of feature importance over time.

3.2.5. Feasibility

The feasibility analysis investigates how the number of available training histories affects the RMSE performance of both models. To assess this, a progressive reduction approach is used: in each iteration, five degradation histories are randomly removed from the training set, and both models are retrained.

Figure 9 illustrates the results of this experiment. The AHSMM demonstrates robust performance, maintaining a relatively stable RMSE until the training set is reduced to just two histories. In contrast, the BiLSTM model shows greater sensitivity to the size of the training data. Its performance begins to degrade noticeably when fewer than 14 training histories are available, highlighting its higher data dependency compared to the AHSMM.

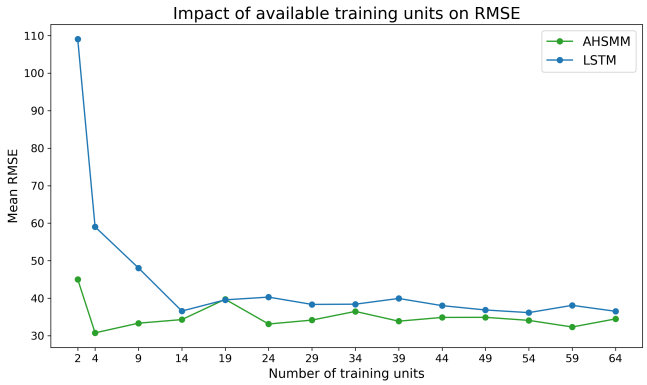

Figure 9. Impact of number of training histories on accuracy error.

4. CONCLUSION AND FUTURE WORK

This paper analyzes data-driven prognostic models through four key characteristics: uncertainty, robustness, interpretability, and feasibility. Prognostic models are broadly categorized as Deep Learning (DL) or stochastic. A case study using the C-MAPSS dataset compared a probabilistic BiLSTM (DL) and an AHSMM (stochastic), highlighting their respective strengths and limitations.

Uncertainty quantification (UQ) remains a major challenge. Stochastic models capture and report uncertainty effectively but often yield wide confidence intervals. DL models tend to be more accurate but often overlook UQ, risking unreliable predictions, though some integrate probabilistic frameworks. This highlights the need to better understand uncertainty sources and data requirements.

Robustness is critical under varying conditions. DL models show promise through domain adaptation but face challenges in industrial settings. Stochastic models like AHSMM adapt better to unseen scenarios. Both performed similarly on RMSE, but BiLSTM struggled with degradation patterns absent from training data, especially near end-of-life, crucial for PHM.

Interpretability is essential in safety-critical domains. AHSMM offers clear, intuitive degradation representations. DL models, including BiLSTM, are often black boxes, though tools like SHAP can partially address this. Balancing interpretability and performance remains challenging.

Feasibility depends on data availability. Techniques like few-shot learning and Bayesian inference help address scarcity. In the case study, AHSMM maintained performance with just two training histories, while BiLSTM's performance declined, emphasizing the need for models effective under limited data.

These differences arise from theoretical foundations. Stochastic models, based on Markovian assumptions and explicit probabilistic frameworks, naturally support UQ, robustness through adaptable degradation modeling, interpretability via state transitions and fewer parameters, and feasibility with limited data. DL models, as flexible deterministic approximators, capture complex nonlinearities but lack inherent uncertainty and interpretability, typically requiring larger datasets and methods like Bayesian extensions to improve robustness, since they map monitoring data directly to RUL rather than modeling degradation.

In summary, no single model excels across all four characteristics. Both DL and stochastic approaches offer valuable strengths but involve trade-offs among accuracy, uncertainty, robustness, interpretability, and feasibility.

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... Acharya, U. R. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, *76*, 243–297.

Alamaniotis, M. (2023). Explainable prognostics method through differential evolved rvr ensemble of relevance vector machines. In *Annual conference of the phm society* (Vol. 15, p. -).

Alcibar, J., Aizpurua, J. I., & Zugasti, E. (2024). Towards a probabilistic fusion approach for robust battery prognostics. *arXiv preprint arXiv:2405.15292*.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. doi: https://doi.org/10.1016/j.inffus.2023.101805

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

Baptista, M., Mishra, M., Henriques, E., & Prendinger, H. (2024, 11). Using explainable artificial intelligence to interpret remaining useful life estimation with gated recurrent unit. *Annual Conference of the PHM Society*, *16*, -. doi: 10.36001/phmconf.2024.v16i1.4124

Caceres, J., Gonzalez, D., Zhou, T., & Droguett, E. L. (2021). A probabilistic bayesian recurrent neural network for remaining useful life prognostics considering

epistemic and aleatory uncertainties. *Structural Control and Health Monitoring*, *28*(10), e2811.

Coble, J., & Hines, J. W. (2009). Identifying optimal prognostic parameters from data: a genetic algorithms approach. In *Annual conference of the phm society* (Vol. 1, p. -).

Costa, N., & Sánchez, L. (2022). Variational encoding approach for interpretable assessment of remaining useful life estimation. *Reliability Engineering & System Safety*, *222*, 108353.

da Costa, P. R. d. O., Akçay, A., Zhang, Y., & Kaymak, U. (2020). Remaining useful lifetime prediction via deep domain adaptation. *Reliability Engineering & System Safety*, *195*, 106682.

Deng, W., Nguyen, K. T., Gogu, C., Medjaher, K., & Morio, J. (2024). Enhancing prognostics for sparse labeled data using advanced contrastive self-supervised learning with downstream integration. *Engineering Applications of Artificial Intelligence*, *138*, 109268.

Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, *31*(2), 105–112.

Ding, P., Jia, M., Ding, Y., Cao, Y., Zhuang, J., & Zhao, X. (2023). Machinery probabilistic few-shot prognostics considering prediction uncertainty. *IEEE/ASME Transactions on Mechatronics*.

Ding, P., Xia, J., Zhao, X., & Jia, M. (2024). Graph structure few-shot prognostics for machinery remaining useful life prediction under variable operating conditions. *Advanced Engineering Informatics*, *60*, 102360.

Eleftheroglou, N., Galanopoulos, G., & Loutas, T. (2024). Similarity learning hidden semi-markov model for adaptive prognostics of composite structures. *Reliability Engineering & System Safety*, *243*, 109808.

Eleftheroglou, N., & Loutas, T. (2016). Fatigue damage diagnostics and prognostics of composites utilizing structural health monitoring data and stochastic processes. *Structural Health Monitoring*, *15*(4), 473–488.

Eleftheroglou, N., Zarouchas, D., & Benedictus, R. (2020). An adaptive probabilistic data-driven methodology for prognosis of the fatigue life of composite structures. *Composite Structures*, *245*, 112386.

Figueroa Barraza, J., López Droguett, E., & Martins, M. R. (2021). Towards interpretable deep learning: a feature selection framework for prognostics and health management using deep neural networks. *Sensors*, *21*(17), 5888.

Folgoc, L. L., Baltatzis, V., Desai, S., Devaraj, A., Ellis, S., Manzanera, O. E. M., ... Glocker, B. (2021). Is mc dropout bayesian? *arXiv preprint arXiv:2110.04286*.

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. arxiv 2019. *arXiv preprint arXiv:1912.02757*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Garreau, D. (2023). Chapter 14 - theoretical analysis of lime. In J. Benois-Pineau, R. Bourqui, D. Petkovic, & G. Quénot (Eds.), *Explainable deep learning ai* (p. 293-316). Academic Press. doi: https://doi.org/10.1016/B978-0-32-396098-4.00020-X

Gebraeel, N., Lei, Y., Li, N., Si, X., & Zio, E. (2023). Prognostics and remaining useful life prediction of machinery: advances, opportunities and challenges. *Journal of Dynamics, Monitoring and Diagnostics*, 1–12.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Huang, C., Bu, S., Lee, H. H., Chan, C. H., Kong, S. W., & Yung, W. K. (2024). Prognostics and health management for predictive maintenance: A review. *Journal of Manufacturing Systems*, *75*, 78–101.

Huang, Z., Xu, Z., Ke, X., Wang, W., & Sun, Y. (2017). Remaining useful life prediction for an adaptive skew-wiener process model. *Mechanical Systems and Signal Processing*, *87*, 294–306.

Kamariotis, A., Tatsis, K., Chatzi, E., Goebel, K., & Straub, D. (2024). A metric for assessing and optimizing data-driven prognostic algorithms for predictive maintenance. *Reliability Engineering & System Safety*, *242*, 109723.

Kobayashi, K., & Alam, S. B. (2024). Explainable, interpretable, and trustworthy ai for an intelligent digital twin: A case study on remaining useful life. *Engineering Applications of Artificial Intelligence*, *129*, 107620.

Kontogiannis, T., Salinas-Camus, M., & Eleftheroglou, N. (2025). Hidden markov models for aviation prognostics. In *Stochastic modeling and statistical methods* (p. 1). Academic Press.

Kundu, R. K., & Hoque, K. A. (2023). Explainable predictive maintenance is not enough: quantifying trust in remaining useful life estimation. In *Annual conference of the phm society* (Vol. 15, p. -).

Li, J., Li, X., & He, D. (2019). Domain adaptation remaining useful life prediction method based on adabn-dcnn. In *2019 prognostics and system health management conference (phm-qingdao)* (pp. 1–6).

Li, N., Wang, M., Lei, Y., Si, X., Yang, B., & Li, X. (2024). A nonparametric degradation modeling method for remaining useful life prediction with fragment data. *Reliability Engineering & System Safety*, *249*, 110224. doi: https://doi.org/10.1016/j.ress.2024.110224

Li, N., Xu, P., Lei, Y., Cai, X., & Kong, D. (2022). A self-data-driven method for remaining useful life prediction of wind turbines considering continuously varying speeds. *Mechanical Systems and Signal Processing*, *165*, 108315. doi: https://doi.org/10.1016/j.ymssp.2021.108315

Li, X., Zhang, W., Ma, H., Luo, Z., & Li, X. (2020). Data alignments in machinery remaining useful life prediction using deep adversarial neural networks. *Knowledge-Based Systems*, *197*, 105843.

Lin, Y.-H., & Li, G.-H. (2022). A bayesian deep learning framework for rul prediction incorporating uncertainty quantification and calibration. *IEEE Transactions on Industrial Informatics*, *18*(10), 7274–7284.

Loutas, T., Eleftheroglou, N., & Zarouchas, D. (2017). A data-driven probabilistic framework towards the in-situ prognostics of fatigue life of composites based on acoustic emission data. *Composite Structures*, *161*, 522–529.

Moghaddass, R., & Zuo, M. J. (2014). An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process. *Reliability Engineering & System Safety*, *124*, 92–104.

Pei, H., Si, X.-S., Hu, C., Li, T., He, C., & Pang, Z. (2022). Bayesian deep-learning-based prognostic model for equipment without label data related to lifetime. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *53*(1), 504–517.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Salinas-Camus, M., & Eleftheroglou, N. (2024). Uncertainty in aircraft turbofan engine prognostics on the c-mapss dataset. In *Phm society european conference* (Vol. 8, pp. 10–10).

Salinas-Camus, M., Goebel, K., & Eleftheroglou, N. (2025). A comprehensive review and evaluation framework for data-driven prognostics: Uncertainty, robustness, interpretability, and feasibility. *Mechanical Systems and Signal Processing*, *237*, 113015. doi: https://doi.org/10.1016/j.ymssp.2025.113015

Serradilla, O., Zugasti, E., Cernuda, C., Aranburu, A., de Okariz, J. R., & Zurutuza, U. (2020). Interpreting remaining useful life estimations combining explainable artificial intelligence and domain knowledge in industrial machinery. In *2020 ieee international conference on fuzzy systems (fuzz-ieee)* (pp. 1–8).

Sharma, J., Mittal, M. L., & Soni, G. (2024). Condition-based maintenance using machine learning and role of interpretability: a review. *International Journal of System Assurance Engineering and Management*, *15*(4), 1345–1360.

Verhagen, W. J., Santos, B. F., Freeman, F., van Kessel, P., Zarouchas, D., Loutas, T., ... Heiets, I. (2023). Condition-based maintenance in aviation: Challenges and opportunities. *Aerospace*, *10*(9), 762.

Vollert, S., & Theissler, A. (2021). Challenges of machine learning-based rul prognosis: A review on nasa's c-mapss data set. In *2021 26th ieee international conference on emerging technologies and factory automation (etfa)* (pp. 1–8).

Xie, F.-Y., Hu, Y.-M., Wu, B., & Wang, Y. (2016). A generalized hidden markov model and its applications in recognition of cutting states. *International Journal of Precision Engineering and Manufacturing*, *17*, 1471–1482.

Zhang, W., Li, X., Ma, H., Luo, Z., & Li, X. (2021). Transfer learning using deep representation regularization in remaining useful life prediction across operating conditions. *Reliability Engineering & System Safety*, *211*, 107556.

Zhu, R., Chen, Y., Peng, W., & Ye, Z.-S. (2022). Bayesian deep-learning for rul prediction: An active learning perspective. *Reliability Engineering & System Safety*, *228*, 108758.

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*. doi: 10.1016/j.ress.2021.108119

BIOGRAPHIES

Mariana Salinas-Camus is a Ph.D. candidate in the iSP group within the Faculty of Aerospace Engineering at TU Delft. She earned her B.Sc. and M.Sc. degrees in Electrical Engineering from Universidad de Chile in 2020 and 2023, respectively, and completed a research internship at NASA Ames Prognostics Center of Excellence during her master's studies. Her current research interests include prognostics, uncertainty quantification, and Bayesian modeling.

Dr. Kai Goebel has spent most of his career investigating different topics for PHM at GE Corporate Research, NASA Ames Research Center, and Palo Alto Research Center (PARC). He has published more than 400 papers, was awarded 24 patents, and has made available numerous public run-to-failure datasets on the NASA Ames data repository.

Dr. Nick Eleftheroglou is an Assistant Professor in the Faculty of Aerospace Engineering at Delft University of Technology and the head of the iSP Group. He received his Diploma in Mechanical and Aeronautics Engineering, cum laude, from the University of Patras, Greece, in 2015, and earned his PhD, cum laude, from TU Delft in October 2020. His research interests lie in the area of prognostics and health management (PHM), developing PHM frameworks with enhanced reliability, robustness, and feasibility for operations and maintenance.