Ensemble of Hybrid Neural Networks to Compensate for Epistemic Uncertainties: A Case Study in System Prognosis

Arinan De Piemonte Dourado (arinan.dourado@louisville.edu)  
University of Louisville  https://orcid.org/0000-0002-0793-9577

Felipe Viana  
University of Central Florida

Research Article

**Keywords:** Physics-informed neural networks, Ensemble learning, Prognosis and health management, Corrosion-fatigue prognosis

**Posted Date:** October 28th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-863306/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Ensemble of hybrid neural networks to compensate for epistemic uncertainties: a case study in system prognosis

Arinan Dourado\textsuperscript{1} and Felipe Viana\textsuperscript{2}\textsuperscript{†}

\textsuperscript{1}Department of Mechanical Engineering, University of Louisville, Louisville, 40292, KY, USA.
\textsuperscript{2}Department of Mechanical and Aerospace Engineering, University of Central Florida, Orlando, 32816, FL, USA.

*Corresponding author(s). E-mail(s): arinan.dourado@louisville.edu;
Contributing authors: viana@ucf.edu;
†These authors contributed equally to this work.

Abstract

In this contribution, a case study considering an unexpected corrosion-fatigue crack propagation issue in an aircraft fleet is used to discuss how to compensate for incomplete knowledge in time dependent responses integration and extrapolation. For the considered application, degradation resulting from mechanical fatigue is well understood and accounted in the damage models. However, the unexpected corrosion effects are not accounted in damage integration, yielding a large discrepancy between predicted and observed crack lengths. To address this epistemic uncertainty in the fleet damage accumulation model, hybrid neural networks cells are formulated; where physics-informed layers address well-understood aspects of the degradation, and data-driven layers are trained to act as correction terms. The considered case study encompasses highly imbalanced data sets with uncertainties acting asynchronously. To improve overall accuracy, ensemble learning techniques are adapted to merge the resulting hybrid neural network cells predictions. Lastly, a heuristic based on optimal ensemble weights is presented to help in the decision-making task of defining safe operation of the fleet. Results show that our proposed approach was capable of compensating for the epistemic uncertainties, and that the proposed heuristic can be used to rank aircraft damage severity, allowing to prioritize aircraft for inspection and/or route reassignment.

Keywords: Physics-informed neural networks, Ensemble learning, Prognosis and health management, Corrosion-fatigue prognosis

1 Introduction

In industrial segments, as civil aviation, in which degradation is intrinsically related to safety, successful implementation of maintenance programs, requires the ability to properly account for unexpected and/or unforeseen factors. There are several published reports in which failure to identify and account for uncertainties introduced during aircraft service life lead to serious incidents (an illustrative example can be found in \textit{NTSB Contributors} (1988)). Hence, prognosis and health management (PHM) analyses (i.e., the ability to model and predict the system degradation state) under these uncertain conditions requires
an understanding of the complex stochastic interactions between operating conditions and system capability. In a way, under such complex and some times unforeseen scenarios, that can lead to large discrepancies between predicted and observed useful life, much of the work resumes on building reliable models for unexpected failure modes and/or machine degradation rates, i.e., models that can compensate for the unaccounted uncertainties effects. Additionally, prognosis models in segments as civil aviation, are used in decision-making scenarios, regarding asset allocation (which aircraft will operate in a given route) or maintenance planning. Such analyzes requires the ability to properly integrate and forecast in time the rate of degradation for quantities of interest (i.e., extrapolate components accumulated damage state), and are crucial for properly managing aircraft fleets. Thus, the ability to identify and compensate for uncertainties effects, especially epistemic uncertainties (i.e., uncertainties due to incomplete or lack of knowledge), are of utmost importance for prognosis analyses.

Prognosis and health management in itself already encompasses several challenges involving an array of distinct aspects with a rich literature devoted to such topics Jardine et al (2006); J.Z.Sikorska et al (2011); An et al (2015); Lei et al (2018); Diez-Olivan et al (2019); Guo et al (2019); Fink et al (2020). Concerning aviation industry applications, prognosis analysis reflects the complexity of the aircraft as a system, with the modeling task usually focusing on critical sub-systems (e.g., aircraft engines, airframe, actuators, landing gear, and avionics). With regards to airframe components, although there might be many reasons for failure, in terms of prognosis, the two most important forms are fatigue and corrosion Findlay and Harrison (2002); Bhaumik et al (2008). Discussions considering crack propagation issue can be found in DuQuesnay et al (2003); Nagaraja et al (2007); Xie et al (2016); Li et al (2017).

A crucial aspect in prognosis and health management is asset remaining useful life (RUL) estimation. Here the goal is to properly characterize the damage state of any given component, and more importantly accurately estimate is future state. Thus, in the task of RUL estimation for prognosis analysis the key challenge is the extrapolation of damage accumulation models under the influence of varying levels of uncertainty. There is a vast literature concerning uncertainty quantification and estimation in PHM analyzes and RUL estimation Baraldi et al (2010); Sankararaman (2015); Dewey et al (2019); Li et al (2021); Caceres et al (2021). In this contribution we focus specifically on the effects of epistemic uncertainty (i.e., incomplete or lack of knowledge) in damage extrapolation for prognosis analysis. A case study considering an unexpected corrosion-fatigue crack propagation issue in a fleet, is used to address the issue of model form uncertainty compensation. For the considered case study, component degradation for low cycle mechanical fatigue is well understood and accounted in the damage accumulation model. However, corrosion effects due to an unexpected exposure to corrosive environments are not accounted by in the damage integration, leading to a large discrepancy between predicted and observed crack length values Additionally, the consider application is plagued by highly imbalanced data sets, with very sparse observations of accumulated damage, i.e., outputs in the form of crack length values are only available for portion of the fleet during inspection campaigns, resulting in very few crack observations, further hindering the task of deriving accurate prediction models.

As illustrated by Jiang et al (2018); Hüllermeier and Waegeman (2021), epistemic uncertainties effects in model predictions are usually consider by means of probabilistic frameworks. Even in contributions where non-probabilistic frameworks are proposed as in Clavreul et al (2013); Tang et al (2017); Zhang et al (2018), in which elements of evidence and fuzzy theories are applied; employed approaches are often still closely related to probability theory. In this contribution we departure from probabilistic frameworks and propose a deterministic approach capable of compensating for incomplete knowledge of the degradation mechanism in the prognosis analysis. Given that damage accumulation can be interpreted as an initial value problem, where the current damage state is estimated by the integration of ordinary differential equations, hybrid recurrent neural network cells are formulated to account for the missing-physics issue in the damage models. In the proposed hybrid recurrent cells, physics-informed layers address well-understood aspects of the damage accumulation, while data-driven layers are trained
to account for the epistemic effects, acting as correction terms. Due to the large imbalance in the data sets for the considered application (i.e., severely sparse label set), in this contribution we opted to adapt ensemble learning techniques to improve the overall accuracy of the final model prediction. The key idea is to merge predictions of hybrid physics-informed neural networks, each derived from distinct physical considerations, to account for the uncertain factors, yielding more reliable predictions.

Several contributions reports the use of ensemble learning methods to increase accuracy and/or reliability and robustness in models predictions. In Che et al (2019) a prognosis model for health assessment of aircraft systems is derived by combining multiple deep learning algorithms. Studies as shown in Hu et al (2012) in which the focus lies on the weighting scheme impacts on the overall remaining useful life prediction; or as in Li et al (2019b,a) and Shi et al (2021) in which different aspects of ensemble learning are evaluated for the remaining useful life estimation of aircraft bearings and engines; illustrates the main capabilities and contributions of ensemble learning on the scope of prognosis and health management. In Guo et al (2020a) ensemble learning is used to address class-imbalance problems, while Guo et al (2020b) and Yousefnezhad et al (2021) employed ensemble learning to increase prediction accuracy and robustness in classification problems. Chou et al (2020) proposed an ensemble learning framework to predict shear strength of reinforced concrete beams. Obtained results illustrated the ability of the proposed frameworks to outperformed other machine learning frameworks potentially serving as a better alternative to help in the design of such structures.

As in the aforementioned contributions, here we employ ensemble methods to improve the accuracy of the considered damage accumulation model, but here the considered damage model is inherently biased due to the partial knowledge of the degradation mechanism acting on the fleet. Hence, each base model on the ensemble needs first to learn how to compensate for the epistemic uncertainty effects in its predictions. Additionally, the adapted ensemble technique can be employed to help in the decision-making task of defining a time window for safe operation of each aircraft based on its forecasted damage and predefined safety thresholds. To do so, in this contribution a heuristic based on optimal ensemble weights is presented.

The remaining of the paper is organized as follows. Section 2 presents the considered numerical experiment and the synthetic fleet data generation. Section 3 details the proposed hybrid physics-informed neural network cells alongside the physical aspects considered in each of these cells. The main aspects concerning ensemble learning used in this contribution are briefly discussed in Section 4. In Section 5, the main results of this contribution are presented and discussed. Finally, Section 6 closes the paper recapitulating salient points and presenting conclusions and future work.

2 Case Study - Unexpected Corrosion-Fatigue in Aircraft Fleet

For the discussions presented in this paper, we considered a scenario in which due to unforeseen air-carrier operational variations, an aircraft fleet is mainly operating in coastal routes. Such conditions significantly increases aircraft exposure to corrosive environments, conditions that are not accounted in the fleet prognosis analysis, leading to an unexpected corrosion-fatigue problem. Root cause analysis on inspected aircraft would potentially identify the corrosion effects, and in the long-term manufacturers would be able to update fleet prognosis models with high-fidelity degradation models accounting for such effects. However, in the meantime, operators would still need to reliably run the fleet, take decisions concerning aircraft route structures, and inspection campaigns. Hence, tools capable of rapidly and accurately compensating for the discrepancies introduced by the unaccounted corrosion effects would be critical. Mathematically, such issues consists of an epistemic uncertainty problem, i.e., uncertainty due to incomplete knowledge of the problem, and in a simplified interpretation can be seen as an unknown bias that is plaguing the prognosis model. Different hybrid physics-informed neural networks will be proposed in the following sections to compensate for such epistemic uncertainty.
To emulate such operational conditions, we used a numerical experiment yielding in a synthetic aircraft fleet. The previously discussed corrosion-fatigue problem is addressed in the synthetic fleet by predicting crack propagation of an Al 2024-T3 alloy on critical points of aircraft wing panels under varying loading and environmental conditions. Hence, panels are subjected to cyclic mechanical loading alongside saline corrosion.

Crack propagation in the fleet is addressed by a cumulative damage model Fatemi and Yang (1998); Frangopol et al (2004) describing the irreversible accumulation of damage throughout the useful life of components. For that matter, fatigue crack propagation is usually modeled through Paris law Paris and Erdogan (1963):

$$\frac{da}{dt} = C(\Delta K_t)^m$$  \hspace{1cm} (1)

where $\Delta K$ is the stress intensity range and $C$ and $m$ are material properties determined through coupon data and many engineering materials have constants documented in handbooks such as MMPDS collaborators (2017).

The Paris law can be modified to take into account the effect of mean stress, expressed in terms of the stress ratio between the minimum and maximum stress levels $R = S_{min} / S_{max}$. We use the Walker model Dowling (2012) to incorporate the stress ratio, that in its discrete form is given by:

$$a_t = a_{t-1} + \frac{C_0}{(1 - R_t)^m(1 - \gamma)}(\Delta K_t)^m$$  \hspace{1cm} (2)

where $C_0$ and $m$ depend of environmental conditions and $\gamma$ depends of material and loading conditions. Stress intensity range, $\Delta K$ comes from engineering models that use information about geometry, crack length, and internal loads (stress amplitude). Assuming that fatigue damage accumulates under mode I loading condition, the stress intensity range $\Delta K_t$ can be expressed as $\Delta K_t = F \Delta S_t \sqrt{a_{t-1}}$, where $F$ is a geometry factor and $\Delta S_t$ are far-field stresses.

Corrosion-fatigue is a very complex phenomenon involving pit nucleation, pit growth, fatigue crack nucleation, short crack growth, transition from short crack to long crack, and long crack growth. Modeling corrosion-fatigue has proven to be challenging Goswami and Hoeppner (1995); Shi and Mahadevan (2001) and it is not our objective to propose a new method (nor defend any preferred method) for it. Instead, we will focus on how a deep neural network based on an Euler integration cell Viana et al (2021) can be designed and trained to compensate for the corrosion effects in crack propagation.

In any given flight, the vast majority of fatigue cycles are purely mechanical. However, take-off, landing, and taxing cycles are subjected to a mix of mechanical loading and saline corrosion. The level of saline corrosion-fatigue depends on the airport and weather related aspects (seasonality). As shown by empirical studies, such as the one found in Menan and Henaf (2010), corrosion effects drastically increase the damage accumulation rate (shifting the $da/dN$ vs $\Delta K$ curve up, when compared to the rates in purely mechanical-fatigue, see Figure 1). To address such effects, in this paper, for the synthetic fleet data generation, we define a corrosivity index $c_t$, which essentially controls the interpolation between the pure air and 3.5% NaCl curves (see Figure 1). Pure air represents a purely mechanical-fatigue condition, while degradation in a saline condition of NaCl 3.5% is the most aggressive corrosive environment here considered. The exact degradation curve will depend of the corrosivity index value. We use the curve for air when $c_t = 0$ (and thus the crack propagation is purely mechanical and a function of the coefficients for pure air shown in Figure 1), while the curve for NaCl at 3.5% is used when $c_t = 1$ (crack propagation is of a corrosion-fatigue condition based then on the coefficients for NaCl shown in Figure 1). Values of $c_t$ in the interval $0 < c_t < 1$ represents the unknown environmental conditions and are related to the unknown bias that the proposed hybrid neural networks are trained to compensate.

The control points of interest are consider to have an initial damage condition that can be represented as an initial crack length value ($a_0 = 0.5$ mm). Here, we arbitrarily considered that the crack length information derived from inspection is initially available for only part of the fleet after 9,000 flights (this value was selected to approximated what would be the second “C” check (see Vieira and Loures (2016)) in the fleet). In this case study, we consider that input conditions (i.e.,
far-field stresses $\Delta S_t$, stress ratio $R_t$, and corrosivity index $c_t$ are fully observed. In real life, $\Delta S_t$ and $R_t$, i.e., loading conditions, could be estimated using engineering analysis with models relating maneuvers to loads. For the analysis presented in this contribution, the corrosivity index $c_t$ is estimated using meteorological data (available in an airport-by-airport basis) by adapting a modified version of the PACER LIME model Summitt and Fink (1980); Roberge et al (2002). Further details on the data generation for this case study are discussed in the Appendix Section.

In the recurrent neural network terminology, different implementations of $f(\cdot)$ are referred to as cells. A cell can be as simple as a single-layer perceptron, but it can also assume sophisticated forms such as the long short-term memory (LSTM) Hochreiter and Schmidhuber (1997) and the gated recurrent unit (GRU) Chung et al (2015). In this paper, we use the Euler integration cell Viana et al (2021). This cell was conceived to implement numerical integration of ordinary differential equations using recurrent neural networks. Nascimento and Viana Nascimento and Viana (2020), Dourado and Viana Dourado and Viana (2020), and Yucesan and Viana Yucesan and Viana (2020) demonstrated that the Euler integration cell can be implemented as a hybrid model where data-driven nodes compensate for model-form uncertainty.

The case study presented in this paper consists of an initial value problem that can be directly integrated as:

$$a_n = a_0 + \sum_{t=1}^{n} \Delta f(x_t, a_{t-1}).$$

where $\Delta f(x_t, a_{t-1})$ combines physics of fatigue damage accumulation and a data-driven node for estimation of corrosion-fatigue damage increment.

At this point one might be tempted to ask why use a hybrid recurrent neural network cell for the damage integration instead of employing well established frameworks as the LSTM or GRU cells given that such cells have reports of successful implementation in seemingly more complex applications? The answer lies in the niche of application we target in this paper. Our application of cumulative damage models involves full observation of input conditions (i.e., loads and environmental parameters), and partial observation of damage (crack length). This scenario leads to highly imbalanced data sets plagued by asynchronous uncertainties (i.e., only few cycles in the damage accumulation are subject to corrosive damage) for training and validation. As shown by Nascimento and Viana Nascimento and Viana (2020) and Yucesan and Viana Yucesan and Viana (2021), under these circumstances established recurrent neural network cells (e.g., LSTM or GRU) do not achieve the same performance as the considered hybrid cells,
unless the number of observations are increase; which is not representative for the considered case study, given that this would imply performing unscheduled inspection campaigns.

### 3.1 General Correction Term: Bias Model

In this study, we assume that fleet prognosis model only considers the failure mechanism of mechanical fatigue. Therefore, damage would be estimated integrating Eq. (2) using $C_0$ and $m$ for a non-corrosive environment (pure air). However, as previously discussed in Section 2 due to operational variations the fleet is exposed to the combined failure mechanism of corrosion-fatigue. Hence, our proposed approach is to train hybrid physics-informed neural networks to model the damage accumulation and compensate for the corrosion effects. While the base damage increment of mechanical fatigue is given by the physics-based model of Eq. (2), here, a data-driven node adjusts the base damage increment to account for the missing physics of corrosion. Thus, the damage accumulation model in this case can be expressed as:

$$a_t = a_{t-1} + \frac{C_0}{(1 - R_t)^{m(1-\gamma)}}(\Delta K_t)^m + g(\Delta S_t, R_t, c_t, a_{t-1}) \tag{5}$$

where $g(\cdot)$ is the unknown functional related to the missing contribution of the corrosion effects in the crack propagation. Here, we use a multi-layer perceptron (MLP) to model corrosion-fatigue damage increment, i.e., $\Delta a_{C,t} = MLP(\Delta S_t, R_t, c_t, a_{t-1})$.

The multi-layer perceptron takes $\Delta S_t, R_t, c_t, a_{t-1}$ as inputs and returns $\Delta a_{C,t}$, which is never observed (only observed output is the accumulated damage in the form of crack length at specific inspection points). Figure 2a illustrates the resulting recurrent neural network cell in which, its physics-informed portion (highlighted in blue) implements Eq. (2), while its data-driven portion (the MLP highlighted in green) is trained to approximate the unknown functional $g(\cdot)$ of Eq. (5). The multi-layer perceptron is adjusting the damage increment outputted by the physics-informed layers so that it compensates for the effects of corrosion. Therefore, it works as a discrepancy-correction term, or simply as a general bias term.

![Fig. 2: Recurrent neural network and Euler integration cell for each proposed base model.](image-url)
3.2 Revising the Inputs for the Bias Term: Bias$_{\Delta K}$ Model

This functional form is a direct review of the general bias cell discussed throughout Section 3.1. The bias estimator present in Section 3.1 consists of a general corrector term of functional form $g(\Delta S_t, R_t, c_t, a_{t-1})$. Due to the underlying functional form of this MLP it acts as a simple general bias term, impacted by any unaccounted effects manifested in its inputs (loading condition and environmental effects). In practice, as stated, the data-driven model detailed in Section 3.1 acts as a “black box”, proposing correction terms without focusing in its “root cause”.

Here, we aim to minimize the contribution of the loading condition in the bias term by revising its inputs, as shown in Eq. (6). This revised functional it is still derived as a bias term, proposing correction values to compensate for the missing-physics in the mechanical fatigue damage increment term ($\Delta a_{F,t}$) as in the previous model:

$$\Delta a_{C,t} = MLP(\Delta K_t, R_t, c_t) \quad (6)$$

where $\Delta a_{C,t}$ is the bias term related to the missing-physics issue that the data-driven portion of the cell (i.e., the MLP) aims to learn during training.

Despite seemly looking like a minor modification to the model detailed in Section 3.1, the physical interpretation, and numerical repercussion are very significant. While in the previous cell the data-driven layer is acting as a generalized bias term for the damage accumulation process, here the estimated bias values are more constrained in form. The revised functional form takes the stress intensity range $\Delta K_t$ as input, instead of observing directly the far-field stresses $\Delta S_t$ values. This means that the contributions of these stresses to the unknown bias term are not directly inferred, but instead will be indirectly manifested in the stress intensity range $\Delta K_t$. Such features may hinder the ability of the proposed cell to identify potential impacts of the loading conditions in the correction terms. The final result potentially is a cell more focused on the environmental effects (represented in form by $c_t$) in the crack propagation. Thus, we expect that for the case study considered, this hybrid cell will yield the most accurate predictions. Figure 3a illustrates the main physical concepts used to derive this functional form, while in Figure 2b the resulting recurrent neural network cell is presented. As in the previous case, physics-informed nodes are highlighted in blue, and implements Eq. (2), while the data-driven portion of the cell is highlighted in green and address the MLP described by Eq. (6).
3.3 Estimating the Unknown Degradation Curve: Log Model

In this section, the functional form described by Eq. (7) is derived under distinct physical assumptions. As in the previously described hybrid cells, it still formulated as a bias correction term for the mechanical fatigue damage increment, but contrary to the previous cells, the bias term ($\Delta a_{C,t}$) is being directly calculated based on estimates of the unknown degradation curve (see Figure 3b). Hence, instead of learning to propose correction terms $\Delta a_{C,t}$, the data driven portion of the hybrid cell focus in providing approximations of equivalent material coefficients $C$ and $m$, so that the unknown degradation curve can be directly estimated, and the contribution of the missing-physics can be calculated as:

$$
\Delta a_{C,t} = \Delta a_{F,t} \times (-1 + 10^{\log MLP_C(R_t,c_t) + MLP_m(R_t,c_t) \times \log \Delta K_t}) \quad (7)
$$

where $\Delta a_{F,t}$ is the mechanical fatigue damage increment given by Eq. (2) considering $C_0$ and $m$ values for pure air, $MLP_C$ and $MLP_m$ are the multi-layer perceptrons tasked with estimating values for the equivalent coefficients $C$ and $m$, respectively, as illustrated in Figure 2c.

This is a significant departure from the previously discussed cells. Here the proposed multi-layer perceptrons observe only the stress ratio ($R_t$) and the corrosivity index ($c_t$) at any given cycle in order to estimate values for the material coefficients. This implies that this model will potentially be heavily influenced by environmental effects. Additionally, given that the data-driven nodes of the cell do not observe any cumulative effect in training, either directly in the form of crack length $a_{t-1}$, or indirectly expressed in the stress intensity range $\Delta K_{F,t}$, its training might be more challenging, given that its gradients in back-propagation will not be a direct function of accumulated errors (expressed in $a_{t-1}$ or $\Delta K_{F,t}$) as in the previous cells. Thus, this cell will most likely be more sensitive to its initialization weights in training. Lastly, despite being more focused on the environmental effects in the unknown bias related to the missing-physics problem, this model is still capable of accounting for any unexpected effects (e.g., loading issues, faulty sensors, etc.) by proposing equivalent material coefficient values capable of compensating for such factors.

3.4 Equivalent Stress Intensity Range: $\Delta K$ Model

Lastly, in this section, we propose a functional form more focused on the loading conditions impacts on the missing-physics problem. As detailed in Eq. (8) in this hybrid model, the data-driven layers no longer act as bias estimators for the mechanical fatigue damage increment but rather aims to compensate for missing-physics effects on the stress intensity range ($\Delta K_t$) by proposing correction terms (bias values) for its value

$$
\Delta K_t = \Delta K_{F,t} + \Delta K_{C,t}, \quad \text{and}
\Delta K_{C,t} = MLP(\Delta S_t, R_t, c_t, a_{t-1}) \quad (8)
$$

where $\Delta K_{F,t}$ is the stress intensity range considering purely mechanical fatigue; and $\Delta K_{C,t}$ is a bias term in the stress intensity value, related to the missing-physics, that the data-driven portion of the cell (i.e. the MLP) aims to learn how to compensate for during training. Figure 3c presents the main physical aspects of this model, while an illustration of the resulting recurrent neural network cell can be seen in Figure 2d.

This is a more extreme departure from the modeling approaches previously discussed. Here the focus is on the stress intensity range rather than on the damage increment itself. This is the least representative model for the considered case study of this paper, given that in the proposed case study the source of the epistemic uncertainty is the missing-physics of corrosion, directly related to environmental effects and not significantly impacted by loading conditions. However, the proposed cell is still capable of yielding accurate predictions for the considered case study by proposing equivalent stress intensity values to compensate for the missing-physics of corrosion. Albeit this is not representative of the actual physics of degradation in the case study, the numerical results might be the same, as long as the hybrid model learns to stipulate a valid equivalency between the corrosion effects and the stress intensity range correction terms.
4 Ensemble Learning: Weighted Average Ensembles

In this section, we focus on discussions concerning aggregation techniques to combine the previously described neural networks predictions and increase the overall accuracy. By adapting the weighted average surrogate procedure detailed in Viana et al. Viana et al (2009), weighted average ensembles can be derived using cross-validation errors, merging thus predictions of these pre-trained neural networks. Hence, a weighted average ensemble intends to take advantage of several neural networks into an ensemble or if it is advantageous to combine previously described base models, namely, Bias, BiasΔK, Log, and ΔK models) aiming to canceling errors in prediction through proper weighting selection in the linear combination of the models and thus yielding better predictions.

Mathematically, this can be summarized as:

\[ a_{	ext{Ensemble}} = \sum_{i=1}^{n} \omega_i \times a_i = \omega^T a \quad \text{and} \]

\[ \sum_{i=1}^{n} \omega_i = 1 \quad \tag{9} \]

where \( a_{	ext{Ensemble}} \) is the predicted crack length by the resulting ensemble, \( \omega_i \) is the weight associated with the \( i \)th base model, and \( a_i \) is the predicted response by the respective base model (Bias, BiasΔK, Log, and ΔK models). As detailed by Viana et al. Viana et al (2009), given a \( C \) matrix of base models cross-validation errors, the optimal weighted ensemble can be found through the minimization problem:

\[ \min_{\omega} \omega^T C \omega, \quad \text{subjected to} \quad \mathbf{1}^T \omega = 1, \quad \tag{10} \]

where \( c_{ij} = \frac{1}{p} e_i^T e_j \), \( p \) is the number of data points in the cross-validation, and \( e_i \) and \( e_j \) are the cross-validation errors of base models \( i \) and \( j \), respectively.

The solution of the optimization detailed in Eq. (10) is obtained through Lagrange multipliers as presented by Eq. (11).

\[ \omega = \frac{C^{-1}}{\mathbf{1}^T C^{-1} \mathbf{1}} \quad \tag{11} \]

Thus, by adopting such procedure to built the neural networks ensemble brings the key advantage of using closed form solution to determine the ensemble weights, not requiring thus, additional computational costs of performing the ensemble weights optimization. In the following sections optimal weights for the considered case study will be presented and discussed. Additionally, a brief discussion concerning the impact of the base models discrepancy (i.e., the variability in the base models prediction trends) on the ensemble accuracy will be addressed.

Lastly, we devise a heuristic metric based on the ensemble weights to guide the decision-making task of defining a time-window for safe operation of the fleet. Here, the goal is to define a time range for prediction extrapolation (fleet prognosis) in which we are confident that the ensemble is still accurate enough to determine aircraft operability. To do so, we considered the ensemble weights as metrics of confidence in each of its base models, i.e., a weight of 90% for instance, indicates that we have a 90% confidence in the respective base model prediction. Hence, a heuristic metric can be derived based on the amount of flights predicted by each base model to achieve a given crack length threshold (\( a^{TH} \)) defined on an operational basis:

\[ \Delta t_{\text{Forecast}}^{\text{Ensemble}} = \sum_{i=1}^{n} \omega_i \times \Delta t_{a \rightarrow a^{TH}}^{i} \quad \tag{12} \]

where \( \Delta t_{\text{Forecast}}^{\text{Ensemble}} \) is the total number of flights predicted by the ensemble to achieve the given threshold (actively defining a time-window for damage forecast and fleet prognosis analysis), \( \Delta t_{a \rightarrow a^{TH}}^{i} \) are base models prediction for the number of flights to achieve the crack length threshold, and \( \omega_i \) are the base models respective weights.

5 Results and Discussions

In this section we will present and discuss results related to the fundamental questions of this paper, mainly, focusing if it is advantageous to combine several neural networks into an ensemble or if is better to simply use the best base model. In other
words, if there is a gain, in prediction terms, of using the proposed ensemble approach. We start by discussing the use of cross-validation analysis to properly rank the base models and guide the selection of the best base model. Then, we shift our discussion to the ensemble responses, in terms of optimal weights and a gain analysis in comparison with the best base model. Lastly, we discuss the use of the ensemble weights to determine a time-window for damage forecast in the fleet.

5.1 Physics-informed Neural Network Design

In order to build the previously detailed hybrid physics-informed neural network cells, the first step is to define the architecture for the multi-layer perceptrons of each cell. Dourado and Viana (2020) presented a discussion on multi-layer perceptron architecture selection based on cross-validation. Here, for simplicity, we will use the architectures presented in Table 1. Training of the recurrent neural network cells were performed using RMSProp as an optimizer, set up with varying values of learning rate (from $10^{-9}$ up to $10^{-3}$) considering 25 epochs. Mean squared error (MSE) loss function was adopted for all the recurrent neural network cells training.

The optimization of recurrent neural network hyperparameters can be challenging and it is conducted through a gradient-descent algorithm. An initial guess for the hyperparameters that is far away from the optimum might cause divergence of the optimization process or at least significantly increase of training time. In this paper, we also used the auxiliary plane approach proposed in Dourado and Viana (2020) to help in the initialization of multi-layer perceptron hyperparameters. This procedure was used only to yield an initial point for the recurrent neural network cells training and it does not interfere with the cells overall performance.

5.2 Cross-Validation Analysis and Model Ranking

In this section, we evaluate the use of cross-validation errors to directly infer the ranking of the base models and guide the selection of the “best” base model in the pool. To carry such an analysis we first need to define a cross-validation procedure from which we will extract the cross-validation errors. We adopted cross-validation procedures as presented in Kohavi et al. (1995). To perform this analysis we applied the following measures: we split the fleet into portions, from which, 20 sample sets containing each 20 aircraft derived from combinations of the same 40 aircraft were defined. On each sample set, a cross-validation analysis, considering a “leave-one-out” strategy was carried. This procedure is repeated until all aircraft in the sample set were removed once.

Figure 4a illustrates the root mean squared error (RMSE) distribution of the cross-validation errors for each base model considering all the sample sets. This cross-validation analysis seems to point out that in the vast majority of the samples the “best” model was the $\text{Bias}_\Delta K$ model. The ranking from the remaining models varies on a sample basis, but considering the average behavior shown in Figure 4a, the cross-validation analysis indicates a ranking in which the best models are the Bias models, with the Log model

| Layer # | neurons/activation function | $\text{Bias}_\Delta K$ model | $\text{Bias}$ model | $\text{Log}$ model $C$ | $\text{Log}$ model $m$ |
|---------|-----------------------------|-------------------------------|---------------------|------------------------|------------------------|
| 0       | 10 / linear                 | 5 / tanh                      |                     |                        |                        |
| 1       | 5 / linear                  | 1 / tanh                      |                     |                        |                        |
| 2       | 1 / sigmoid                 |                               |                     |                        |                        |
| Parameters* | 131                         | 38                           |                     |                        |                        |

| Layer # | neurons/activation function | $\Delta K$ model |
|---------|-----------------------------|-------------------|
| 0       | 5 / tanh                    |                   |
| 1       | 1 / relu                    |                   |
| 2       | 1 / sigmoid                 |                   |
| Parameters* | 51                          |                   |

* differences in the total number of parameters are due to differences in the number of inputs.

---

1 https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/RMSprop
being the third in the pool, while the $\Delta K$ model being ranked as last. This ranking is in agreement with the expected behavior for the considered corrosion-fatigue case-study given the aspects of each base model discussed in Section 3. Figure 4b brings a comparison of the ranking considering the cross-validation analysis and the true ranking on each sample. In all the considered sample sets the $\text{Bias}_{\Delta K}$ model was the most accurate model, while the $\Delta K$ model was the least accurate. Albeit not identifying the true ranking on all the sample sets, the considered cross-validation procedure can be safely used to segregate the base models, and identify the overall ranking of the base models.

![Cross-validation errors for each base model.](image)

(a) Cross-validation errors for each base model.

![Model ranking comparison.](image)

(b) Model ranking comparison.

**Fig. 4:** Model selection based on cross-validation errors: (a) - Errors across sample sets (considered crack lengths are in the range of approximately 1.0 to 6.0 mm across the samples sets); (b) - Base models true ranking (derived from synthetic fleet actual prediction error and unknown in practical applications) against model ranking based on cross-validation errors.

### 5.3 Merging the Predictions: Ensemble Analysis

In this section we present and discuss the main findings considering the previously discussed weighted average ensemble approach (see Section 4) to merge the base neural networks predictions. Figure 5a illustrates a possible set of weights for the neural network ensemble considering the cross-validation errors of a given sample set. For simplicity presented weights were rounded up. As we can see the derived weights highly favor the $\text{Bias}_{\Delta K}$ model and hardly differentiate the remaining base models. This a direct reflection of the behavior illustrated in the cross-validation error distribution shown in Figure 4a. A possible explanation for the ensemble weights to slightly favor the $\Delta K$ model over the $\text{Log}$ model in the example shown in Figure 5a, despite the $\Delta K$ model clearly overestimation of larger crack lengths, is most likely due to the fact that the best model in this case (the $\text{Bias}_{\Delta K}$ model) slightly underestimate such cracks. It is worth highlighting that Figure 5a presents a single example of ensemble weights. Many other combinations of weights can be found considering the other sample sets in the cross-validation analysis.

The results illustrated by Figure 5a helps demonstrating the task of merging the base models predictions. However, in Figure 5a we can observe that the ensemble responses are very close to the predictions of the best base model ($\text{Bias}_{\Delta K}$ model) in the pool (due to the high weight value associated with such model). Albeit such behavior may not be observed in other sample sets, it raises an important question. There is any gain of using such ensemble predictions instead of simply using the best base model? Or simply putting what is the advantage of using such ensembles in comparison to just using the best base model predictions? To address such questions, we perform a gain analysis for the ensemble predictions based on the metric define as:

$$ gain(\%) = 100 \times \left( \frac{RMSE_{\text{Best}} - RMSE_{\text{Ensemble}}}{RMSE_{\text{Best}}} \right) $$

where $RMSE_{\text{Best}}$ is always related to the true best base model on a sample set basis. Therefore it changes with the considered data set and it is an idealization, in the sense that it
would not be easily define in practical applications given that is derived from actual errors (based on true crack length for all aircraft in the sample). $RMSE_{Ensemble}$ is related to the ensemble prediction errors.

The results of this gain analysis is shown in Figure 5b with varying number of base models in the ensemble. Except for two outliers we can see that when the number of base models is equal to 1 (i.e., the ensemble is composed only of the best base model) the described metric is zero. This illustrates once again how the ranking based on the cross-validation errors analysis can safely and accurately identify the best base model in the pool ($RMSE_{Best}$).

When the number of base models in the ensemble is larger than 1, the defined metric can either be positive or negative. Negative values indicate that for the considered weight combination in the ensemble, the predictions will be worse than the ones generated by the best base model, meaning thus that, at this condition, it does not pay off to combine the base models. Positive values indicate that the ensemble predictions are more accurate than the ones generated by the best base model and thus it is worth it to combine the model predictions. In Figure 5b we can observe that as the number of base models in the ensemble increases, the gain associated with such ensembles also increases, meaning that the ensemble accuracy surpasses the accuracy of best base model. This is an indicative that the ensembles gain is related to its base model diversity.

To better illustrate the role that the base models diversity plays on the ensemble performance consider the scenario presented in Figure 6b. In this example the resulting ensemble presents a negative gain value, indicating that its predictions are worst than the best base model in the pool. But a closer inspection on the base models behavior for this scenario indicates an underlying bias from the base models, i.e., for this scenario all base models have an underestimation trend. In other words, this scenario represents a lack of diversity condition for the ensemble, and at such conditions the resulting ensemble will invariably yield worsen predictions than the best base model. Figure 6a illustrates a scenario in which a clear spread concerning the base model predictions can be observed, i.e., there is a “disagreement” amongst the base model predictions. In such conditions the inherent diversity in the base models prediction will help yielding an ensemble that can outperform the best base model. These results further collaborates with the observation that even relatively “poor” base models actively contributes to the ensemble accuracy, and suggests that there is no gain in removing the “worst” base model from the ensemble.

(a) Example of an ensemble prediction. Weights derived considering the cross-validation errors in a given sample set.

(b) Evaluation on the percentage gain of using an ensemble over using solely the predictions of the best base model, considering all 20 sample sets.

Fig. 5: Ensemble predictions for a given sample and ensemble gain evaluation.

Lastly, we focus our discussions on the performance of the proposed metric to determine “confiability” of damage forecast. Figure 7a illustrates the key ideas of proposed heuristic metric for a given aircraft in a given sample set. Each base model yields different predictions for the number of flights to achieve the arbitrary crack length threshold ($a^{TH}$) of 20 mm, depending of its underlying conservativeness and accuracy. The proposed metric weight the base models estimates considering the ensemble optimal weights into
an average prediction. It is worth mentioning an interesting contribution of the proposed heuristic metric. The ensemble prediction for the number of flights to achieve the defined threshold will not coincide with the metric estimate. This is due the accumulated effect of the base models over or underestimation in the ensemble forecast. The ensemble prediction is given by the linear combination of the base models estimates for crack length on every loading cycle, while the heuristic is the linear combination of a single value (i.e., the number of flights until the threshold). This observation highlights the value of the proposed heuristic. In Figure 7b we present the comparison between the predicted number of flights to achieve the predefined crack threshold given by the proposed metric, and the actual number of flights required to achieve this crack value. The comparison considers only the aircraft that surpassed the defined threshold within 13,000 flights or less in a given sample set (i.e., we considered here 4,000 flights of damage forecast). The proposed metric predictions are relatively close to the actual number of flights, with observed differences in the range of few hundreds flights. Additionally, for the results shown in Figure 7 we can see that the proposed metric predictions has an underlying conservatism concerning the actual number of flights. Albeit this observed conservatism is highly related to the underlying behavior of the base models in the ensemble, in a safety related application as in aviation, a given level of conservatism in prognosis is certainly desired. The presented results are a strong indicative of the accuracy of the proposed heuristic, indicating that the proposed approach could safely be adopted in fleet prognosis analysis, helping ranking aircraft damage in the fleet and helping prioritizing inspection.

5.4 Replication of Results

Simulations were conducted using a laptop configured with an Intel Core i7-7820HQ CPU at 2.90GHz, 32GB of RAM, and NVIDIA Quadro M620 graphical processing unit running Windows 10. Our implementation is all done in TensorFlow using the Python application programming interface (version 2.0). In case this paper is accepted, the data and codes needed for replication of our results will be made publicly available on GitHub under the MIT License.

6 Summary and Closing Remarks

Motivated by the advances in hybrid modeling frameworks, capable of merging elements of first principles reduced-order models and machine learning, in this paper, we focus on the development of weighted ensembles of hybrid neural networks capable of minimizing the impact of epistemic uncertainties in model prediction and extrapolation. Despite the advances in uncertainty propagation, especially for the field of system prognosis and health management, epistemic uncertainty estimations are still a challenging task. To illustrate the key concepts of the proposed formulation, a numerical study considering an
Fig. 7: Sample results for damage forecast evaluation considering the proposed heuristic. (a) illustrates the concepts used to define Eq (12). (b) presents the predicted number of flights in a given sample set considering the proposed metric. The observed difference in the number of flights to achieve the considered threshold were relatively small when comparing to the actual number (all in the range of few hundreds of flights).

unexpected exposure of an aircraft fleet to highly corrosive environments leading to an unaccounted corrosion-fatigue crack propagation issue was presented. The fleet prognosis model only accounts for mechanical fatigue in its damage accumulation analysis, ultimately yielding in an epistemic uncertainty due to the missing physics of the corrosive damage.

Firstly, we presented distinct approaches to derived hybrid physics-informed recurrent neural network cells to accounted for the epistemic uncertainties effects on damage accumulation. Each neural network cell was defined based on specific physical interpretations of the corrosion-fatigue problem, proposing each, unique ways to account for the missing physics problem. The major achievement of such hybrid models is the ability to use data-driven layers to compensate for the missing physics in reduced-order models and accurately estimate damage accumulation. Additionally, the physics-informed layers reduce the need for larger data sets found in purely data-driven approaches. Presented results indicate that despite varying levels of conservativeness in the hybrid models, all proposed hybrid cells are able to capture the overall trend of fleet crack propagation, and in its own way compensate for the missing physics of corrosion.

Then, we evaluated the use of weighed average ensembles, derived from the linear combination of the proposed hybrid cells. Ensemble weights were defined based on cross-validation analysis, and the key idea of the proposed approach is to take advantage of existing pre-trained neural networks aiming to canceling errors in prediction through proper weighting selection, ultimately yielding better predictions. Discussions concerning the advantage of using the proposed neural network ensemble over the best base model were addressed. Results here presented shown the impact of the base models diversity in the ensemble gain. Here results point out to the conclusion that increasing the number of base models in the ensemble, associated to an increment in the “discordance” of the base models predictions, significantly increases the ensemble gain in accuracy. Lastly, we devised an heuristic metric based on the ensemble optimal weights to define a time-window for safe operation of the fleet. As illustrated by the results presented in this contribution, the proposed heuristic is capable of accurately defining a valid time-window for forecast, indicating that the proposed metric could safely be adopted in fleet prognosis analysis.

Despite the promising results, we acknowledge that the proposed approach should be further developed. Concerning future directions we suggest: evaluating a possible convergence of the ensemble accuracy concerning the number of base models; investigate the ensemble performance when multiple sources of uncertainty plague the damage accumulation process; extend the proposed framework to consider complex combined failure modes simultaneously acting on the fleet; and evaluate if the proposed neural network ensemble is capable of automatically learn to identify the root cause of the unexpected observations and properly detect such unforeseen factors.
Declarations

- **Funding:** This work was supported by the University of Central Florida (UCF).
- **Conflict of interest/Competing interests:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Ethics approval:** N/A
- **Consent to participate:** N/A
- **Consent for publication:** N/A
- **Availability of data and materials:** N/A
- **Code availability:** In case this paper is accepted, the data and codes needed for replication of our results will be made publicly available on GitHub under the MIT License.
- **Authors’ contributions:**
  - Arinan Dourado: Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing, Visualization.
  - Felipe Viana: Conceptualization, Methodology, Validation, Software, Formal Analysis, Investigation, Writing, Supervision, Funding Acquisition.

Appendix A Synthetic Aircraft Fleet Data

A hypothetical control point on the underside of an aircraft wing, as illustrated by Figure A1a, is considered here. For simplicity, we consider that aircraft in the fleet can fly the 10 different flight types detailed in Table A1. Based on the load magnitude each loading cycle of the flight can be translated into far-field stress ($\Delta S = S_{max} - S_{min}$) and stress ratio ($R = S_{min}/S_{max}$) values. Each flight type is characterized by a load frequency distribution. Therefore, flights present different severity levels in terms of mechanical loads. In order to balance out the exposure of the fleet to the severe flights, we considered the 10 arbitrarily designed mission mixes. In our study, each aircraft in the fleet is assigned to only one specific mission mix. Within a given mission mix, flights types are assigned to the aircraft randomly following the probability distribution detailed in Table A2.

In this paper, we assumed that corrosive cycles represent, on average, 5% of the total number of cycles in one mission (loads of type E and F are related to take-off and landing, respectively, and may be subjected to corrosion-fatigue). We then proceed to penalize this cycles by accumulating damage using curves that are between air and NaCl at 3.5% (see Figure 1). In our model, the exact curve will depend of a corrosivity index $c_t$. While the actual relationship between the corrosivity index and the $C_0$ and $m$ constants is

![Control point on the underside of the aircraft wing.](image_a1a)

![Modified PACER LIME model.](image_a1b)

**Fig. A1:** Synthetic aircraft fleet data generation information.

| Flight type | Load frequency |
|-------------|----------------|
| A           | 2              |
| B           | 2              |
| C           | 1              |
| D           | 3              |
| E           | 5              |
| F           | 13             |
| G           | 2              |
| H           | 12             |
| I           | 22             |
| J           | 1              |
| K           | 12             |
| L           | 2              |
| M           | 12             |
| N           | 2              |
| O           | 12             |
| P           | 2              |
| Q           | 12             |
| R           | 2              |
| S           | 12             |
| T           | 2              |
| U           | 12             |
| V           | 2              |
| W           | 12             |
| X           | 2              |
| Y           | 12             |
| Z           | 2              |

**Table A1:** Flight type load matrix distribution and related normalized minimum and maximum stresses (MPa) (adapted from Jonge et al (1973)).

| Flight type | Load frequency |
|-------------|----------------|
| A           | 0.07           |
| B           | 2.26           |
| C           | 4.46           |
| D           | 6.65           |
| E           | 8.84           |
| F           | 11.0           |
| G           | 28.3           |
| H           | 26.0           |
| I           | 23.8           |
| J           | 21.7           |
| K           | 19.5           |
| L           | 17.3           |

| Flight type | Load frequency |
|-------------|----------------|
| A           | 0.07           |
| B           | 2.26           |
| C           | 4.46           |
| D           | 6.65           |
| E           | 8.84           |
| F           | 11.0           |
| G           | 28.3           |
| H           | 26.0           |
| I           | 23.8           |
| J           | 21.7           |
| K           | 19.5           |
| L           | 17.3           |

of cycles in one mission (loads of type E and F are related to take-off and landing, respectively, and may be subjected to corrosion-fatigue). We then proceed to penalize this cycles by accumulating damage using curves that are between air and NaCl at 3.5% (see Figure 1). In our model, the exact curve will depend of a corrosivity index $c_t$. While the actual relationship between the corrosivity index and the $C_0$ and $m$ constants is
unknown; here, we arbitrarily used the following equation to generate synthetic data:

\[
C_0 = -2.275 \times 10^{-8} (c_t)^2 + 4.505 \times 10^{-8} c_t + 1.132 \times 10^{-10},
\]

and

\[
m = 2.046(c_t)^2 - 4.052c_t + 3.859. \tag{A1}
\]

Several factors need to be considered while determining the corrosivity index \(c_t\). Here, we use a modified version of the PACER LIME model presented in Figure A1b. In this case study we considered corrosion index values coming from a truncated normal distribution, (i.e., \(0 < c_t < 1\)), with scale \(\sigma = 0.05\), and location (\(\mu_{c_t}\)) defined by the modified PACER LIME model severity levels (see Figure A1b). In our modified model, this levels are: “AA”, very severe corrosivity level (\(c_t = 1\)); “A”, severe corrosivity level (\(\mu_{c_t} = 0.844\); “B”, moderate corrosivity level (\(\mu_{c_t} = 0.608\)); and “C”, mild corrosivity level (\(\mu_{c_t} = 0.361\)). On each flight mission two airports must be associated to a flight. To help structure this step in the synthetic simulation, we arbitrarily defined 10 routes. Hence, at each flight mission an aircraft is randomly assigned to a given route, and therefore, to specific corrosivity indexes.

Following all the previously discussed steps, the fleet simulation can be summarized as: an aircraft begins the route structure and the severity levels presented in Figure A1b), and the mission mix configuration (see Table A2); the specific flight type for each mission is then randomly assigned following the mission mix probability distribution (see Table A2); based on the selected flight type, loads (i.e., \(\Delta S\) and \(R\) values) are applied following the distributions detailed in Table A1.

### References

An D, Kim NH, Choi JH (2015) Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. Reliability Engineering & System Safety 133:223–236. https://doi.org/10.1016/j.ress.2014.09.014

Baraldi P, Popescu IC, Zio E (2010) Methods of uncertainty analysis in prognostics. International Journal of Performability Engineering 6(4):303. [URL](https://doi.org/10.23940/ijpe.10.4.p303.mag)

Bhaumik S, Sujata M, Venkataswamy M (2008) Fatigue failure of aircraft components. Engineering Failure Analysis 15(6):675 – 694. [URL](https://doi.org/10.1016/j.engfailanal.2007.10.001)

Caceres J, Gonzalez D, Zhou T, et al (2021) A probabilistic bayesian recurrent neural network for remaining useful life prognostics considering epistemic and aleatory uncertainties. Structural Control and Health Monitoring p e2811. [URL](https://doi.org/10.1002/stc.2811)

### Table A2: Mission mix configuration (flight type probability per mission).

| Mix | I     | II    | III   | IV    |
|-----|-------|-------|-------|-------|
| 1   | 5.0 × 10⁻⁵ | 7.5 × 10⁻⁵ | 0.001 | 2.4 × 10⁻³ |
| 2   | 2.5 × 10⁻⁵ | 2.5 × 10⁻⁵ | 1.2 × 10⁻³ | 1.5 × 10⁻³ |
| 3   | 0      | 0      | 1.3 × 10⁻³ | 2.3 × 10⁻³ |
| 4   | 0      | 5.0 × 10⁻⁵ | 9.5 × 10⁻⁴ | 2.5 × 10⁻³ |
| 5   | 0      | 0      | 1.2 × 10⁻³ | 2.3 × 10⁻³ |
| 6   | 1.0 × 10⁻⁴ | 2.5 × 10⁻⁵ | 7.5 × 10⁻⁴ | 2.3 × 10⁻³ |
| 7   | 0      | 0      | 0      | 2.3 × 10⁻³ |
| 8   | 2.5 × 10⁻⁵ | 0      | 0.001 | 2.5 × 10⁻³ |
| 9   | 0      | 7.5 × 10⁻⁵ | 7.5 × 10⁻⁴ | 7.7 × 10⁻³ |
| 10  | 7.5 × 10⁻⁵ | 0      | 7.5 × 10⁻⁴ | 2.4 × 10⁻³ |

| Mix | V     | VI    | VII   | VIII  |
|-----|-------|-------|-------|-------|
| 1   | 0.01  | 0.010 | 0.095 | 0.105 |
| 2   | 6.75 × 10⁻³ | 0.025 | 0.035 | 0.155 |
| 3   | 0.016 | 0.025 | 0.025 | 0.105 |
| 4   | 8.5 × 10⁻³ | 0.013 | 0.095 | 0.055 |
| 5   | 0.006 | 0.015 | 0.045 | 0.105 |
| 6   | 6.13 × 10⁻³ | 0.015 | 0.045 | 0.08  |
| 7   | 7.3 × 10⁻³ | 0.015 | 0.070 | 0.055 |
| 8   | 0.001 | 0.02  | 0.02  | 0.12  |
| 9   | 0.001 | 0.025 | 0.025 | 0.055 |
| 10  | 6.25 × 10⁻³ | 0.005 | 0.05  | 0.08  |

| Mix | IX    | X     |
|-----|-------|-------|
| 1   | 0.2475 | 0.5278 |
| 2   | 0.273 | 0.5025 |
| 3   | 0.273 | 0.5530 |
| 4   | 0.323 | 0.5030 |
| 5   | 0.2975 | 0.5278 |
| 6   | 0.3975 | 0.4528 |
| 7   | 0.3225 | 0.5278 |
| 8   | 0.285 | 0.5528 |
| 9   | 0.2225 | 0.6528 |
| 10  | 0.2725 | 0.5828 |
Che C, Huawei, Wang, et al (2019) Combining multiple deep learning algorithms for prognostic and health management of aircraft. Aerospace Science and Technology 94. https://doi.org/10.1016/j.ast.2019.105423, URL https://doi.org/10.1016/j.ast.2019.105423

Chou JS, Nguyen TK, Pham AD, et al (2020) Shear strength prediction of reinforced concrete beams by baseline, ensemble, and hybrid machine learning models. Soft Computing 24(5):3393–3411. https://doi.org/https://doi.org/10.1007/s00500-019-04103-2

Chung J, Gulcehre C, Cho K, et al (2015) Gated feedback recurrent neural networks. In: Bach F, Blei D (eds) Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 37. PMLR, Lille, France, pp 2067–2075, URL http://proceedings.mlr.press/v37/chung15.html

Clavreul J, Guyonnet D, Tonini D, et al (2013) Stochastic and epistemic uncertainty propagation in lca. The International Journal of Life Cycle Assessment 18(7):1393–1403. https://doi.org/https://doi.org/10.1007/s11367-013-0572-6

Dewey HH, DeVries DR, Hyde SR (2019) Uncertainty quantification in prognostic health management systems. In: 2019 IEEE Aerospace Conference, pp 1–13, https://doi.org/https://doi.org/10.1109/AERO.2019.8741821

Diez-Olivan A, Del Ser J, Galar D, et al (2019) Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. Information Fusion 50:92 – 111. https://doi.org/https://doi.org/10.1016/j.inf fus.2018.10.005, URL http://www.sciencedirect.com/science/article/pii/S1566253518304706

Dourado A, Viana FAC (2020) Physics-informed neural networks for missing physics estimation in cumulative damage models: a case study in corrosion fatigue. ASME Journal of Computing and Information Science in Engineering 20(6):061,007 (10 pages). https://doi.org/10.1115/1.4047173, URL https://doi.org/10.1115/1.4047173

Dowling NE (2012) Mechanical Behavior of Materials: Engineering Methods for Deformation, Fracture, and Fatigue. Pearson

DuQuesnay DL, Underhill PR, Britt HJ (2003) Fatigue crack growth from corrosion damage in 7075-T651 aluminum alloy under aircraft loading. International Journal of Fatigue 25(5):371–377

Fatemi A, Yang L (1998) Cumulative fatigue damage and life prediction theories: a survey of the state of the art for homogeneous materials. International Journal of Fatigue 20(1):9 – 34. https://doi.org/10.1016/S0142-1123(97)00081-9

Findlay S, Harrison N (2002) Why aircraft fail. Materials Today 5(11):18 – 25. https://doi.org/https://doi.org/10.1016/S1369-7021(02)01138-0, URL http://www.sciencedirect.com/science/article/pii/S1369702102011380

Fink O, Wang Q, Svensén M, et al (2020) Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications of Artificial Intelligence 92:103,678. https://doi.org/https://doi.org/10.1016/j.engappai.2020.103678, URL http://www.sciencedirect.com/science/article/pii/S0952197620301184

Frangopol DM, Kallen MJ, van Noortwijk JM (2004) Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. Progress in Structural Engineering and Materials 6(4):197–212. https://doi.org/10.1002/pse.180

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, URL http://www.deeplearningbook.org

Goswami TK, Hoeppner DW (1995) Pitting corrosion fatigue of structural materials. In: Structural Integrity in Aging Aircraft. ASME, New York, p 47

Guo H, Zhou J, Wu Ca (2020a) Ensemble learning via constraint projection and undersampling
technique for class-imbalance problem. Soft Computing 24(7):4711–4727. https://doi.org/https://doi.org/10.1007/s00500-019-04501-6

Guo J, Li Z, Li M (2019) A review on prognostics methods for engineering systems. IEEE Transactions on Reliability 69(3):1110–1129. https://doi.org/10.1109/TR.2019.2957965, URL https://doi.org/10.1109/TR.2019.2957965

Guo Y, Wang X, Xiao P, et al (2020b) An ensemble learning framework for convolutional neural network based on multiple classifiers. Soft Computing 24(5):3727–3735. https://doi.org/https://doi.org/10.1007/s00500-019-04141-w

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu C, Youn BD, Wang P, et al (2012) Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliability Engineering & System Safety 103:120–135. https://doi.org/10.1016/j.ress.2012.03.008, URL https://doi.org/10.1016/j.ress.2012.03.008

Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning 110(3):457–506. https://doi.org/10.1007/s10994-021-05946-3

Jardine AKS, Lin D, Banjevic D (2006) A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing 20(7):1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012, URL http://www.sciencedirect.com/science/article/pii/S0888327005001512

Jiang C, Zheng J, Han X (2018) Probability-interval hybrid uncertainty analysis for structures with both aleatory and epistemic uncertainties: a review. Structural and Multidisciplinary Optimization 57(6):2485–2502. https://doi.org/https://doi.org/10.1007/s00158-017-1864-4

Jonge JBD, Schutz D, Lowak H, et al (1973) A standardized load sequence for flight simulation tests on transport aircraft wing structures. Report from National Aerospace Laboratory, Amsterdam, NLR TR 73029U

J.Z.Sikorska, M.Hodkiewicz, L.Ma (2011) Prognostic modelling options for remaining useful life estimation by industry. Mechanical Systems and Signal Processing 25(5):1803–1836. https://doi.org/10.1016/j.jsyspop.2010.11.018

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., IJCAI’95, pp 1137–1143

Lei Y, Li N, Guo L, et al (2018) Machinery health prognostics: A systematic review from data acquisition to rul prediction. Mechanical systems and signal processing 104:799–834. https://doi.org/10.1016/j.ymssp.2017.11.016, URL https://doi.org/10.1016/j.ymssp.2017.11.016

Li C, Mahadevan S, Ling Y, et al (2017) Dynamic Bayesian network for aircraft wing health monitoring digital twin. AIAA Journal 55(3):930–941. https://doi.org/10.2514/1.J055201

Li G, Yang L, Lee CG, et al (2021) A bayesian deep learning rul framework integrating epistemic and aleatoric uncertainties. https://doi.org/https://doi.org/10.1109/TIE.2020.3009593

Li Z, Goebel K, Wu D (2019a) Degradation modeling and remaining useful life prediction of aircraft engines using ensemble learning. Journal of Engineering for Gas Turbines and Power 141(4). https://doi.org/10.1115/1.4041674, URL https://doi.org/10.1115/1.4041674

Li Z, Wu D, Hu C, et al (2019b) An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. Reliability Engineering & System Safety 184:110–122. https://doi.org/10.1016/j.ress.2017.12.016, URL https://doi.org/10.1016/j.ress.2017.12.016
Menan F, Henaft G (2010) Synergistic action of fatigue and corrosion during crack growth in the 2024 aluminum alloy. Procedia Engineering 2:1441–1450. https://doi.org/10.1016/j.proeng.2010.03.156

MMPDS collaborators (2017) MMPDS - 12: Metallic Materials Properties Development and Standardization

Nagaraja I, Subhasis S, Robert M, et al (2007) Aircraft life management using crack initiation and crack growth models-p-3c aircraft experience. International Journal of Fatigue 29(9-11):1584–1607. https://doi.org/10.1016/j.ijfatigue.2007.03.017

Nascimento RG, Viana FAC (2020) Cumulative damage modeling with recurrent neural networks. AIAA Journal 58(12):5459–5471. https://doi.org/10.2514/1.J059250, URL https://arc.aiaa.org/doi/10.2514/1.J059250

NTSB Contributors (1988) Aircraft Accident Report, Aloha Airlines Flight 243, Boeing 737-100,N73711, Near Maui, Hawaii, April 28, 1988. Aircraft accident report, National Transportation Safety Board, URL https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR8903.pdf

Paris P, Erdogan F (1963) A critical analysis of crack propagation laws. Journal of Basic Engineering 85(4):528–533. https://doi.org/10.1115/1.3656900

Roberge P, Klassen R, Haberecht P (2002) Atmospheric corrosivity modeling – a review. Materials & Design 23(3):321 – 330. https://doi.org/10.1016/S0261-3069(01)00051-6, URL http://www.sciencedirect.com/science/article/pii/S0261306901000516

Sankararaman S (2015) Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. Mechanical Systems and Signal Processing 52-53:228–247. https://doi.org/10.1016/j.ymssp.2014.05.029, URL https://doi.org/10.1016/j.ymssp.2014.05.029

Shi J, Yu T, Goebel K, et al (2021) Remaining useful life prediction of bearings using ensemble learning: The impact of diversity in base learners and features. Journal of Computing and Information Science in Engineering 21(2):021,004. https://doi.org/10.1115/1.4048215, URL https://doi.org/10.1115/1.4048215

Shi P, Mahadevan S (2001) Damage tolerance approach for probabilistic pitting corrosion fatigue life prediction. Engineering Fracture Mechanics 68:1493–1507. https://doi.org/10.1016/S0013-7944(01)00041-8

Summitt R, Fink FC (1980) PACER LIME: An environmental corrosion severity classification system. Technical Report AFWAL-TR-80-4102, Air Force Wright Aeronautical Laboratories

Tang H, Li D, Li J, et al (2017) Epistemic uncertainty quantification in metal fatigue crack growth analysis using evidence theory. International Journal of Fatigue 99:163–174. https://doi.org/https://doi.org/10.1016/j.ijfatigue.2017.03.004, URL https://www.sciencedirect.com/science/article/pii/S0142112317300919

Viana FAC, Haftka RT, Steffen V (2009) Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Structural and Multidisciplinary Optimization 39(4):439–457. https://doi.org/10.1007/s00158-008-0338-0, URL http://dx.doi.org/10.1007/s00158-008-0338-0

Viana FAC, Nascimento RG, Dourado A, et al (2021) Estimating model inadequacy in ordinary differential equations with physics-informed neural networks. Computers and Structures 245:106,458. https://doi.org/10.1016/j.compstruc.2020.106458, URL https://www.sciencedirect.com/science/article/pii/S0045794920302613

Vieira DR, Loures PL (2016) Maintenance, repair and overhaul (mro) fundamentals and strategies: An aeronautical industry overview. International Journal of Computer Applications 135(12)
Xie L, Yang Y, Zhou Z, et al (2016) Dynamic neural modeling of fatigue crack growth process in ductile alloys. Information Sciences 364-365:167–183. https://doi.org/https://doi.org/10.1016/j.ins.2016.05.010, URL https://www.sciencedirect.com/science/article/pii/S0020025516303346

Yousefnejad M, Hamidzadeh J, Aliannejadi M (2021) Ensemble classification for intrusion detection via feature extraction based on deep learning. Soft Computing pp 1–17. https://doi.org/https://doi.org/10.1007/s00500-021-06067-8

Yucesan YA, Viana FAC (2020) A physics-informed neural network for wind turbine main bearing fatigue. International Journal of Prognostics and Health Management 11(1):17 pages. https://doi.org/10.36001/ijphm.2020.v11i1.2594, URL https://papers.phmsociety.org/index.php/ijphm/article/view/2594

Yucesan YA, Viana FAC (2021) Hybrid physics-informed neural networks for main bearing fatigue prognosis with visual grease inspection. Computers in Industry 125:103,386. https://doi.org/10.1016/j.compind.2020.103386, URL https://www.sciencedirect.com/science/article/pii/S0166361520306205

Zhang Q, Kang R, Wen M (2018) A new method of level-2 uncertainty analysis in risk assessment based on uncertainty theory. Soft Computing 22(17):5867–5877. https://doi.org/https://doi.org/10.1007/s00500-018-3337-0