An End-to-End Pipeline for Uncertainty Quantification and Remaining Useful Life Estimation: An Application on Aircraft Engines

Marios Kefalas[1], Bas van Stein[2], Mitra Baratchi[3], Asteris Apostolidis[4], and Thomas Bäck[5]

[1,2,3,5] *LIACS, Leiden University, Leiden, 2333 CA, The Netherlands*
{*m.kefalas, b.van.stein, m.baratchi, t.h.w.baeck*}*@liacs.leidenuniv.nl*

[4] *Faculty of Technology, Amsterdam University of Applied Science, Amsterdam, 1097 DZ, The Netherlands*
a.apostolidis@hva.com

ABSTRACT

Estimating the remaining useful life (RUL) of an asset lies at the heart of prognostics and health management (PHM) of many operations-critical industries such as aviation. Modern methods of RUL estimation adopt techniques from deep learning (DL). However, most of these contemporary techniques deliver only single-point estimates for the RUL without reporting on the confidence of the prediction. This practice usually provides overly confident predictions that can have severe consequences in operational disruptions or even safety. To address this issue, we propose a technique for uncertainty quantification (UQ) based on Bayesian deep learning (BDL). The hyperparameters of the framework are tuned using a novel bi-objective Bayesian optimization method with objectives the predictive performance and predictive uncertainty. The method also integrates the data pre-processing steps into the hyperparameter optimization (HPO) stage, models the RUL as a Weibull distribution, and returns the survival curves of the monitored assets to allow informed decision-making. We validate this method on the widely used C-MAPSS dataset against a single-objective HPO baseline that aggregates the two objectives through the harmonic mean (HM). We demonstrate the existence of trade-offs between the predictive performance and the predictive uncertainty and observe that the bi-objective HPO returns a larger number of hyperparameter configurations compared to the single-objective baseline. Furthermore, we see that with the proposed approach, it is possible to configure models for RUL estimation that exhibit better or comparable performance to the single-objective baseline when validated on the test sets.

1. INTRODUCTION

Prognostics and health management (PHM) is a research area with multiple methodologies and functions as a *decision support tool* that aims at minimizing maintenance costs and predicting when a failure could occur by the assessment, prognosis, diagnosis, and health management of engineered systems (Nguyen et al., 2019). The core of PHM is failure prognostics. Failure prognostics refers specifically to the phase involved with predicting future behavior and the system's useful lifetime left in terms of current operating state and the scheduling of required maintenance actions to maintain system health (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006). This useful lifetime left is often called the remaining useful life (RUL) (Nguyen et al., 2019) and is defined as the length from the current time and operating state to the end of the useful life (Si, Wang, Hu, & Zhou, 2011). The notice of pending equipment failure allows for sufficient lead-time so that necessary decisions, personnel, equipment, and spare parts can be organized and deployed, thus minimizing equipment downtime and repair costs. By leveraging RUL estimation[1], industries, such as aerospace, maritime, and energy, can improve maintenance schedules to avoid catastrophic failures and consequently save lives and costs (Zhang, Lim, Qin, & Tan, 2017). The industry has to also assure that its asset utilization is optimum by guaranteeing a timely - but not premature - maintenance. Furthermore, this practice promotes sustainability as the use of spare parts is optimum and no useful life is wasted.

The estimation of the RUL can be done in various ways. *Model-based*, *data-driven* and *hybrid* methods are the most prominent approaches (Nguyen et al., 2019), and in general all methods make some use of the sensor data of the equipment and/or maintenance history. Model-based methods (or

[1] In this work we will be using the terms RUL *prediction* and RUL *estimation* interchangeably, unless otherwise stated.

physics-based methods) rely on an established mathematical model of the system in question and, as a result, call for a thorough understanding of the system's physics and processes. This can be prohibitively costly in terms of time and money due to the amount of time and domain expertise needed to develop and fine-tune such models[2]. On the other hand, data-driven methods are relatively easier to develop as they do not need (a lot of) expert knowledge to develop the model, rendering them domain-agnostic or easily transferable between domains. They can require, however, large amounts of data. Lastly, hybrid (or fusion) methods leverage the advantages of the two previous methods while minimizing their limitations. The previous groups of methods showcase that data-driven approaches are, in general, available to a broader audience due to their domain-agnostic nature, allowing universal applicability, and also because of the plethora of tools that are developed.

Data-driven approaches either fall under the category of classic machine learning (ML) algorithms (such as random forests (RF)) (Zhang et al., 2017; Sateesh Babu, Zhao, & Li, 2016a) or the more recently proposed deep neural networks (DNNs) (Hsu & Jiang, 2018; Listou Ellefsen, Bjørlykhaug, Æsøy, Ushakov, & Zhang, 2019; Zheng, Ristovski, Farahat, & Gupta, 2017). In both cases, though, the estimation of the RUL is a challenging problem. The remaining useful life is not merely a target variable that can be predicted from sensor measurements, but it is a variable that needs to be inferred from a longer trend of degradation patterns and when those begin to occur. In this view, and due to the advances in the general field of artificial intelligence (AI), deep learning (DL) and DNNs have proven to be a successful candidate to the RUL estimation task (Lei et al., 2018; Benker, Furtner, Semm, & Zaeh, 2021; Kefalas, Baratchi, Apostolidis, van den Herik, & Bäck, 2021; Caceres, Gonzalez, Zhou, & Droguett, 2021; Peng, Ye, & Chen, 2020; B. Wang, Lei, Yan, Li, & Guo, 2020). One significant advantage of DNNs lies in their ability to learn features from raw data automatically and extract patterns that can enhance the RUL estimation accuracy (Benker et al., 2021; B. Wang et al., 2020). DNNs owe their success to their representational power and their capacity to learn sets of hierarchical features from simpler features due to their deep, multilayer architectures (Goodfellow, Yoshua Bengio, & Aaron Courville, 2016). However, most of the state-of-the-art DL approaches used in prognostics provide mainly point estimates to their RUL predictions (Peng et al., 2020; Caceres et al., 2021; Biggio, Wieland, Chao, Kastanis, & Fink, 2021). This is because DNNs do not inherently quantify the uncertainty associated with their predictions but instead treat their weights and biases as deterministic values. These predictions, though, are uncertain since they are prone to noise and wrong model inference (see Section 4.4). Specifically, there are two sources of uncertainty, namely *epistemic* (or model) uncertainty and *aleatory* (or data) uncertainty (Hüllermeier & Waegeman, 2021). The former occurs due to inadequate knowledge, data, and representational capacity of the model and the latter due to the inherent uncertainty of the data distribution (Caceres et al., 2021; Abdar et al., 2021). Additionally, from the nature of epistemic uncertainty we can see that it is a *reducible* part of the (total) uncertainty of a modeling process, as it can be reduced on the basis of additional information. On the contrary, aleatory uncertainty is an *irreducible* part of the (total) uncertainty, due to the inherently random effects in the data-generating process (Hüllermeier & Waegeman, 2021). Most problems in engineering involve both sources of uncertainties. However, it may be difficult to distinguish whether a particular uncertainty should be put in the aleatory category or the epistemic category, in the modeling phase (Kiureghian & Ditlevsen, 2009).

The lack of a measure of uncertainty, however, can lead to overly confident decisions (Caceres et al., 2021; Gal & Ghahramani, 2016). When it comes, for example, to cost-critical or safety-critical applications, it is necessary to know how much confidence a DL method has on its prognostic results and even more so when it comes to the RUL estimation (Peng et al., 2020; Biggio et al., 2021; Benker et al., 2021; Caceres et al., 2021). In addition, even though DNNs output predictive probabilities (e.g., image classification), these probabilities are falsely interpreted as model confidence (Gal & Ghahramani, 2016). For example, the probability of the softmax on the final layer of a neural network (NN). will not reflect if the network has knowledge of the input (see also adversarial examples (Szegedy et al., 2014)). Additionally, decision-making based on a single-point estimate is error-prone and leaves no room for the decision-maker to make an actionable choice (Peng et al., 2020). When such an uncertainty estimate is available (see also Section 2) it is often the case that end-users and decision-makers need to choose by lacking broader information, such as distribution of predictions or other statistics that can assist the logistics further.

Furthermore, the end-user or researcher is faced with a multitude of decisions around the hyperparameters of the preprocessing of the data (e.g., label construction for RUL data) and of the learning algorithm (e.g., the number of layers in a DNN). Hyperparameters are not learnt but have to be set a-priori, and they have a large impact on the predictive performance of a method but also uncertainty. On top of that, there can be hyperparameter configurations that allow low prediction error but have (relatively) large uncertainty and vice versa. In such scenarios, where trade-offs exist, it is vital to move towards a more *user-centric* approach, where the end-user can decide which hyperparameter configuration to adopt based on the criticality of the task. As such, hyperparameters need to be considered carefully both in terms of model accuracy and uncertainty estimates.

[2]Model-based methods do not require (a lot of) historical data for their development, making them the only option for the development of models for new systems.

The aforementioned statements motivate our main research question: Can we propose an automated framework for configuring RUL prediction models which are highly accurate and have less estimation uncertainty?

More specifically, our contributions are as follows:

1. We automatically optimize the hyperparameters of the Bayesian deep learning (BDL) model through Bayesian multi-objective optimization, jointly minimizing the RUL prediction error and the combined aleatory and epistemic uncertainties of the estimations. The reasoning behind this is that in certain tasks, there can be conflicts between these two objectives, as we briefly mentioned previously.
2. Together with the model hyperparameters, we further optimize the hyperparameters which are specific to the task of RUL estimation (the RUL label construction, see also Section 4), which is known to have an effect on the algorithmic performance (Sateesh Babu, Zhao, & Li, 2016b). We provide thus, a thorough, end-to-end approach that can further assist researchers and end-users for offline RUL estimation.
3. We adopt a *user-centric* approach that allows the user to estimate the RUL based on the model output, as it promotes a more interpretable RUL decision. We demonstrate how survival curves can provide the end-user with information regarding the RUL and its confidence.
4. We evaluate our multi-objective hyperparameter optimization (HPO) approach against a single objective HPO by taking the harmonic mean (HM) of the objectives. Our approach is validated on two subsets of the widely used C-MAPSS dataset (A. Saxena & K. Goebel, 2008).

The rest of the paper is organized as follows. In Section 2, we present related work in this field and in Section 3, we formally define the problem of the RUL estimation. In Section 4, the proposed method and its modules are introduced and in Section 5 we present the dataset used and discuss the experimental results. Finally, in Section 6 we conclude and discuss the limitations of our framework and suggest future work.

2. Related Work

The field of PHM has been widely credited in the past years with numerous contributions from researchers. Academic interest, industrial applications, as well as the scientific challenge of developing methods to forecast a failure, have been the driving forces. While model-based prognostic methods, such as Kalman filters and their variants (Govaers, 2019; Kalman, 1960), take into account the modeling and data uncertainty, only a few studies in the data-driven domain address this matter, despite its importance (Biggio et al., 2021). Touching upon the previous statement, in this section, we will present related work in the context of uncertainty quantification (UQ) for the RUL estimation, attending only to data-driven approaches. From the traditional ML methods, only Gaussian process regression (GPR) (Rasmussen & Williams, 2006) (also known as Kriging) addresses UQ. GPR is a stochastic interpolation method where unseen locations of a stochastic process are estimated as a linear function of observed values. It can further be understood as a form of Bayesian Inference (BI). Specifically, GPR places a Gaussian prior over the functions that could have generated the observed data. Using Bayes's theorem by combining the Gaussian prior and the Gaussian likelihood function (for tractability), we get the predictive distribution for a new value. However, GPR might not be the optimal model for some data, e.g., the data does not come from a Gaussian process, or the dimensionality is high. Furthermore, the data generating the predictions are not learned automatically as in DL but need proper pre-processing (e.g., feature extraction), and another downside of GPR is the GPR variance, of which is known that it can be over-optimistic (den Hertog, Kleijnen, & Siem, 2006).

In this view, from the data-driven approaches, we will only review recent work that adopted a DL solution. We made this decision because, as also mentioned in Section 1, DL is becoming prominent in data-driven prognostics, as well as there has recently been a lot of attention on UQ for DL (Gal & Ghahramani, 2016; Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015; Osband, Aslanides, & Cassirer, 2018; Abdar et al., 2021). This collection is by no means exhaustive. We refer the interested reader to (Nguyen et al., 2019) and (Krishna & Baghaei, 2019) for a more thorough overview of related work on PHM.

Epistemic Uncertainty The work by Peng et al. (Peng et al., 2020) is a recent data-driven example of UQ in prognostics. The authors present a DL approach from a Bayesian viewpoint to address the confidence of their RUL predictions and implement the Bayesian approximation using Monte Carlo Dropout (MC Dropout) (Gal & Ghahramani, 2016) (see also Section 4.4). Kraus et al. (Kraus & Feuerriegel, 2019) dealt with epistemic uncertainty in prognostics using variational inference (VI) (see also Section 4.4) and combine DL with notions from survival analysis to increase the *intepretability* of the estimation. In the same domain, Wang et al. (B. Wang et al., 2020) used MC Dropout to estimate the epistemic uncertainty of a recurrent convolutional neural network (RCNN) for the RUL estimation. However, none of the previous studies touched upon aleatory uncertainty.

Aleatory Uncertainty Zhao et al. (Zhao, Wu, Wong, Sun, & Yan, 2020), addressed the aleatory uncertainty by using a deep convolutional neural network (DCNN) through a shortened version of the ResNet (He, Zhang, Ren, & Sun, 2015) and assumed that the target RUL values follow a Gaussian distribution with parameters μ and σ being the network's outputs. They also adopted a non-parametric approach by combining the predicted RUL from the network with quantile regression, predicting this way multiple RUL at different quantile levels.

However, this approach did not take into account epistemic uncertainty and, to the extent of our knowledge, there was no HPO.

Epistemic and Aleatory Uncertainties Caceres et al. considered in (Caceres et al., 2021) both epistemic and aleatory uncertainties. They used an explicit form of VI to account for the epistemic uncertainty and addressed the aleatory uncertainty by a probabilistic output layer parameterized by a Gaussian distribution and further performed HPO through grid search. In the same manner, Kim et al. (Kim & Liu, 2021) and Li et al. (G. Li, Yang, Lee, Wang, & Rong, 2021) designed RUL frameworks by taking into account the effects of both epistemic and aleatory uncertainties. They both used MC dropout to address the epistemic uncertainties. Kim et al. (Kim & Liu, 2021) addressed the aleatory uncertainty by a probabilistic output layer parameterized by a Gaussian distribution and assumed a monotonically decreasing relationship between the aleatory uncertainty and RUL and further performed HPO on the number of hidden layers amongst other hyperparameters. Li et al. (G. Li et al., 2021) modeled aleatory uncertainty by a probabilistic output layer following various types of lifetime distributions (Weibull, Gaussian, and Logistic). Benker et al. (Benker et al., 2021) adopted a Bayesian neural network and addressed both uncertainties as well, but took into account the aleatory uncertainty post-training. They further quantified the epistemic uncertainty using a Hamiltonian Monte Carlo method, a more efficient variant of the Markov Chain Monte Carlo (MCMC) methods in high dimensional spaces.

These recent studies have made a great contribution to the field of data-driven prognostics by proposing methods to account for and quantify the uncertainty of their predictions. Nonetheless, there remain perspectives to consider. In more detail, most of the literature reviewed ((Zhao et al., 2020; G. Li et al., 2021; Benker et al., 2021)) did not state any form of HPO and those that did ((Peng et al., 2020; Caceres et al., 2021; Kim & Liu, 2021; Biggio et al., 2021)), did not optimize necessary hyperparameters in the pre-processing stage and used less efficient HPO techniques (e.g., grid search). What is more, the reviewed methods that perform some form of HPO used *only* the RUL prediction error as the only criterion to guide the HPO, as opposed to also taking into account the epistemic and aleatory uncertainties. Lastly, in our literature review, we did not come across any methods that allow the end-user to make an informed RUL prediction based on information output by the model.

3. Problem Definition

The RUL of an asset or system is defined as the length from the current time and operating state to the end of the useful life (Si et al., 2011). Because the adjective *useful* is subjective, the previous definition can be extended to the time when the extent of deviation or degradation of the performance from its expected normal operating conditions exceeds a *pre-defined* threshold (Saxena, Goebel, Simon, & Eklund, 2008), when the system needs to be repaired or retracted. Based on this, we can define the RUL at time $t \in \mathbb{R}_{\geq 0}$ as:

$$RUL(t, \mathbf{D}_t) = \inf\{s \in \mathbb{R}_{\geq 0} : s \geq t \wedge \mathbb{1}_{\mathbb{S}^{\complement}}(CI(s, \mathbf{D}_t))\} - t \ , \tag{1}$$

where *inf* represents the infimum of a set and $\mathbb{1}$ is the indicator function. $\mathbb{S}$ is a user-defined system operating envelope. The operating envelope is a collection of boundary limits that put the integrity of an asset at risk when exceeded. *CI* represents a user-specified condition index, which monitors if the asset has exceeded it's operating constraints. In this case, the *CI* lies in the complement of $\mathbb{S}$ ($\mathbb{S}^{\complement}$), which indicates that the system must be repaired or maintained.

The time t denotes the time at which the prediction needs to be performed. $\mathbf{D}_t$ represents the data generated by an asset used for the RUL prediction of that asset. Most commonly $\mathbf{D}_t$ is sensor measurements recorded in time (e.g., pressure, temperature) accompanied by event labels (e.g., times-to-failure), up until time t. In principle, though, $\mathbf{D}_t$ can be any type of data, structured or not, that can facilitate the estimation.

The quantity $\inf\{s \in \mathbb{R}_{\geq 0} : s \geq t \wedge \mathbb{1}_{\mathbb{S}^{\complement}}(CI(s, \mathbf{D}_t))\}$ in Equation 1 can also be referred to as the *end-of-life* (EoL), to mark that the system's "life", based on user-defined criteria, has come to an end. Ultimately the estimation of RUL amounts to the approximation of the EoL. We should note that the EoL does not necessarily mean that the system has gone through a catastrophic failure but might operate sub-optimally according to user-defined criteria.

Finally, from a *data-driven* perspective, the estimation of the RUL of an asset involves creating a model which is trained on data from the same type of assets. Let U be the set of training data. Each instance $u \in U$ is presented as a multivariate time-series of sensor readings $\boldsymbol{X}_u = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{T(u)}]^T \in \mathbb{R}^{m \times T(u)}$, with $T(u)$ time-steps where the last time-step corresponds to the end-of-life (EoL) of the unit u. Each point $\boldsymbol{x}_t \in \mathbb{R}^m$ is an m-dimensional vector corresponding to readings from m sensors at time t.

4. Proposed Method

Our method works by training a Bayesian deep learning model on training data U presented in the form of multivariate time-series. The steps of our method are summarized as:

1. Data pre-processing by removing any redundant signals, normalizing the remaining sensor values and performing a sliding window transformation.
2. Target-RUL construction to allow supervised learning.
3. Modeling using a BDL model and taking into account the uncertainty of the predictions.
4. Hyperparameter optimization of the hyperparameters of steps 1,2, and 3.

4.1. Pre-Processing

Sensor selection is an initial step of pre-processing multivariate time-series data. It involves filtering the available data from sensor measurements which, for example, either do not exhibit any correlation with the target or have strong correlations with other sensors. In the latter case, we usually discard some of the correlated features. Furthermore, even if no correlation is present, but the sensors do not exhibit any variation, these features can often be discarded as they do not add any valuable information. What is more, having a large number of sensors is not always beneficial for training models as it increases the chance of overfitting.

Pre-processing also involves normalizing the available data to mitigate any effect that different ranges of values or large deviations can have in the subsequent learning phase. Two of the most often used normalization methods are Z-normalization and Min-max normalization:

- Z-normalization (or standardization): This normalization transforms the data into having 0 mean and unit variance as: $x' = (x - \mu)/\sigma$;
- Min-max normalization (or rescaling): This normalization maps the range of the data into $[0, 1]$ or more generally into $[a, b]$ as: $x' = a + \frac{(x - \min(S))(b-a)}{\max(S) - \min(S)}$,

where S is a feature (e.g., a sensor), x, x' are the value and the transformed value of the feature S, and μ, σ are the mean and standard deviation of S, respectively. In addition, a, b are the lower and upper bounds of the projection, and $\min(S)$, $\max(S)$, are the minimum value and maximum value of S, respectively. Normalization is applied on every sensor/feature independently.

As a next step, for each $\boldsymbol{X_u}$, we perform a sliding window transformation with a sequence of length w (window size), in order to enclose the inputs into multidimensional sequential data, which are to be considered as one sample. This transformation allows one to increase the number of training data, standardize the sample input lengths, and accelerate model training (Caceres et al., 2021). *For this work, the window size w is treated as a pre-processing hyperparameter.*

4.2. Target-RUL Construction

We would like to tackle this problem as a regression problem. However, one of the main challenges of RUL estimation is the lack of ground-truth values (Sateesh Babu et al., 2016b). In most cases, the only available data are the data from the sensor measurements. However, these data are not labeled with any information regarding the RUL, such as the times-to-failure. The latter is essential for the training procedure as it carries decisive information that will allow the learner to uncover rules that estimate the RUL given sensor measurements. There are two popular ways to create these labels, namely *linear* and *piece-wise linear* methods (Sateesh Babu et al., 2016b). The former interprets the RUL in the strictest sense, as time to failure. Thus, every time-step is mapped to a value equal to $EoL - t$, where t is the current time-step. This approach, however, implies that the health of the system degrades linearly with usage (Sateesh Babu et al., 2016b). The latter reflects the fact that initially the degradation is negligible, and after a specific point in time, it becomes more evident (see Figure 1 for an example). The point after which the RUL degrades linearly is called the *reflection point* (Hsu & Jiang, 2018). This, way we can construct an RUL curve for each $u \in U$, by mapping each rolling window to the RUL *at the end of that window*. *For this work, the type of label creation method and the reflection point are treated as pre-processing hyperparameters.*

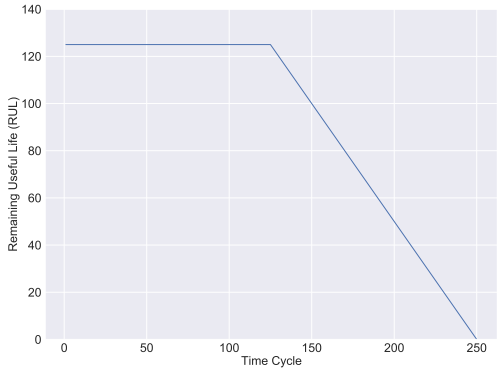

Figure 1. Toy example of a piece-wise linear RUL target function. The reflection point is at time cycle 125. Adapted from (Kefalas et al., 2021).

4.3. Modeling

As mentioned in Section 1, amongst the data-driven methods employed for prognostics DNNs have proven to be good candidates due to their representational power (Lei et al., 2018; Benker et al., 2021; Kefalas et al., 2021; Caceres et al., 2021; Peng et al., 2020; B. Wang et al., 2020). In general, shallow learning methods are not designed for large-scale datasets and, more importantly, need extensive feature engineering efforts (Zhou, Droguett, Mosleh, & Chan, 2021). In this view, we decided to employ DL to address the RUL estimation problem. As this task is based on sequential data (multivariate time-series), we decided to use recurrent layers and specifically gated recurrent unit (GRU) layers as the model base due to their lower complexity and similarly good performance in modeling long dependencies (G. Li et al., 2021), when compared to long short-term memory (LSTM) layers.

4.4. Uncertainty Quantification

As discussed briefly in Section 1 predictions made by neural networks are inherently uncertain, as they are prone to noise and/or wrong model inference. At the same time, however,

NNs treat their weights and biases as deterministic values. This results in NNs being overly confident, even when they should *not* be. In general, there are two sources of uncertainty. In the context of NN, the epistemic and aleatory uncertainties can be considered by putting a prior on model parameters or the outputs. The latter means assuming that the model outputs follow a specific distribution, such as Weibull. The former can be addressed by treating the weights and biases (we jointly note them as W) of the network as random variables, defining a prior over them, and then using Bayesian inference to learn the posterior distributions of the network's weights (Peng et al., 2020; Zhou et al., 2021; Caceres et al., 2021) as:

$$p(W|X,Y) = \frac{p(Y|X,W)p(W)}{p(X,Y)}, \tag{2}$$

where X, Y are the training data and their labels, respectively. The posterior distribution on the network's parameters is, however, computationally intractable even for NNs of any practical size, as the number of parameters is very large and the functional form of a NN does not allow for exact integration (Blundell et al., 2015; Gal & Ghahramani, 2016; Caceres et al., 2021). Moreover, the denominator in Equation 2 is unavailable in closed form or requires exponential time to compute (Blei, Kucukelbir, & McAuliffe, 2017).

A large part of ongoing research is focused on approximating such posterior distributions (Biggio et al., 2021). Amongst these, prominent methods are Markov Chain Monte Carlo (MCMC) methods and its variants and variational inference (VI) (Biggio et al., 2021; Zhou et al., 2021; Blei et al., 2017; Caceres et al., 2021). The former, generally, converge slowly and are computationally expensive for large datasets or complex models. Instead, VI solves the same problem by using optimization techniques rather than sampling methods like MCMC (Blei et al., 2017). Specifically, VI sidesteps the difficulty mentioned above altogether by defining an approximate *variational* distribution $q(W)$ from a distributional family $\mathbb{D}$, that is the best approximation to the exact posterior $p(W|X,Y)$, with respect to the Kullback-Leibler (KL) divergence. This means that,

$$q(W) = argmin_{q(W)\in\mathbb{D}} KL(q(W)||p(W|X,Y)), \tag{3}$$

where $KL(q(W)||p(W|X,Y))$ is defined as:

$$KL(q(W)||p(W|X,Y)) = \mathbb{E}_{q(W)}\left[\log\frac{q(W)}{p(W|X,Y)}\right] \tag{4}$$

However, because Equation 3 is intractable[3] VI maximizes instead what is called the evidence lower bound (ELBO), which is defined as:

$$ELBO(q) = \mathbb{E}\left[\log(p(X,Y|W)\right] - KL(q(W)||p(W)) \tag{5}$$

[3]See (Blei et al., 2017) page 6 for details.

In turn, though, exactly maximizing Equation 5 is computationally prohibitive. To address this, VI can be divided into methods that implicitly use model uncertainties, such as MC Dropout (Gal & Ghahramani, 2016) and methods that explicitly model weight parameters as probability distributions such as Bayes-by-Backprop (Blundell et al., 2015; Caceres et al., 2021; Zhou et al., 2021).

In this work, we have decided to use MC Dropout to model the **epistemic uncertainty** due to its simplicity, scalability, and computational efficiency compared to other Bayesian deep learning approaches (Gal & Ghahramani, 2016; Kim & Liu, 2021). It is implemented through gradient-based learning methods and stochastic regularization techniques, which are widely available in existing DL libraries (Peng et al., 2020). MC Dropout is, in essence, regular dropout applied at both training and inference steps. The addition of dropout between every layer can switch off some portion of neurons in each layer and generate random predictions as samples from a probability distribution that is considered equivalent to performing approximate VI. In more detail, MC Dropout showed that by choosing a specific form of an approximate distribution q, as a distribution over matrices whose columns are randomly set to zero, the VI in a NN can be interpreted as performing one forward pass through the NN with dropout. For more details on MC Dropout, see (Gal & Ghahramani, 2016) and the accompanying appendix.

We should note here that there is a current debate as to the validity of MC Dropout being Bayesian (Caceres et al., 2021; Zhou et al., 2021; Osband et al., 2018). In (Osband et al., 2018), Osband et al. in highlighted that a shortcoming of MC Dropout is that the dropout rate does not depend on the data, which translates into the fact that employing dropout for posterior approximation cannot say anything about a set of data being observed once or more times. This, of course, can have significant implications in support of reliable uncertainty quantification and consequently deserves attention. As this work was mainly devoted to the usage of bi-objective HPO and user-centric approach, we have decided to address this *highly relevant but challenging issue* in future work.

Finally, in order to model the **aleatory uncertainty**, inspired by (Martinsson, 2016), we further assume that the RUL values follow a Weibull distribution, the reason being that Weibull is extensively employed in survival and reliability analysis to model times-to-failure. Moreover, it is simple, but also expressive, being able to take various forms, such as the exponential distribution (G. Li et al., 2021). The probability density function (PDF) of the 2-parameter Weibull that we used is defined as: $f(x) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{b-1}e^{-(x/\alpha)^{\beta}}$, for $x \geq 0, \alpha, \beta \in (0, +\infty)$, where α is the scale parameter and β the shape parameter of the distribution.

In this view and to adopt a user-centric approach for the RUL estimation (3rd contribution), the output layer of the DNN

(see Section 4.3) will output the parameters of the Weibull distribution, α, β. This is a more *user-centric* approach, as for a sample input (e.g., a sequence of sensor values), the end-user is presented with the parameters that govern the distribution of the times-to-failure. This allows for more informative and interpretable decision-making in subsequent steps. The end-user can decide himself what statistics or percentiles (e.g., the mean-time-to-failure (MTTF)) to use as the point estimate of the RUL and the overall knowledge of the distribution of failure times can allow decision-makers to reason if the results are plausible or not. This contrasts with most methods that return a point-estimate to the end-user.

4.5. Hyperparameter Optimization

The optimization of hyperparameters enhances the performance of a machine learning algorithm, and thus, HPO is considered an important step in developing AI and ML frameworks.

Various methods and algorithms are available for HPO, such as grid search (GS), random search (RS), evolutionary algorithms (EA), and Bayesian optimization (BO) (Feurer & Hutter, 2019). In this study, a bi-objective variant of a state-of-the-art BO algorithm, namely *Mixed-integer Parallel Efficient Global Optimization* (MIP-EGO), is chosen due to its efficiency for optimizing expensive problems (Stein, Wang, & Back, 2019). MIP-EGO is based on Efficient Global Optimization, also known as Bayesian Optimization (BO). The algorithm uses random forests (RFs) models to handle mixed integer data and mixed integer evolution strategies (MIES) as internal optimizer. The bi-objective variant of MIP-EGO uses the S-metric hyper-volume (see also Section 5.3) improvement infill criterion to select new candidate solutions.

In order to perform the HPO of the Bayesian deep learning and the problem-specific pre-processing hyperparameters by jointly optimizing the prediction error and uncertainty address (1st and 2nd contributions), MIP-EGO is set to determine the hyperparameter values that *minimize simultaneously* the point-wise root mean squared error (RMSE) and the uncertainty by optimizing the bi-objective function described in Algorithm 1. In more detail, MIP-EGO will evaluate different configurations h_p by preprocessing the data and training a DNN (lines 1 and 2). In lines $3-19$ the trained network is used to make predictions on each sample of the validation set (size m) by multiple passes R which output different α, β at each pass using MC Dropout (see Section 4.4). To determine the RUL estimate for an input sample, we calculated the median of the predicted αs ($\bar{a}$) and the median of the predicted βs ($\bar{b}$) (line 11) and used the mean-time-to-failure (MTTF) of the Weibull distribution with parameters the calculated medians (line 15). The choice of the MTTF was to reduce the selection bias to any statistic and the choice of median to counteract effects of possible outliers. Of course, any other statistic could be used here. The mean-time-to-failure is defined as: $MTTF(\alpha, \beta) = \alpha\Gamma(1 + 1/\beta)$, where Γ is the gamma function. For the over all point-wise performance, f_1, we calculated the RMSE between the predicted RUL (over all the instances) and the ground truth values (line 18). To determine the uncertainty for an input sample, we calculated the standard deviation of the predicted αs ($\hat{a}$) and the standard deviation of the predicted βs ($\hat{b}$) (line 13) and averaged the two values. For the overall uncertainty f_2, we calculated the average over all the uncertainties (line 19).

Algorithm 1: Bi-objective Function

Data: X, V, hp, R # Training data, validation data, hyperparameter configuration, sample size
Result: f_1, f_2 # RMSE, uncertainty
1 $X', V', Y_{X'}, Y_{V'} \leftarrow$ Pre_processing(X, V, hp); # Data pre-processing and RUL creation for the training and validation data (see Sections 4.1 and 4.2)
2 $M \leftarrow$ DNN$(X', V', Y_{X'}, Y_{V'}, hp)$; # Model training
3 $m \leftarrow |V'|$; $RUL \leftarrow \emptyset$; $Var \leftarrow \emptyset$;
4 **for** $i \leftarrow 1$ **to** m **do**
5 $A \leftarrow \emptyset$; $B \leftarrow \emptyset$;
6 **for** $j \leftarrow 1$ **to** R **do**
7 $a, b \leftarrow M(V_i)$;
8 # Predicting using trained DNN through MC Dropout (see Section 4.4)
9 $A \leftarrow A \cup a$; $B \leftarrow B \cup b$;
10 **end**
11 $\bar{a} \leftarrow median(A)$; $\bar{b} \leftarrow median(B)$;
12 # Median values of A and B
13 $\hat{a} \leftarrow std(A)$; $\hat{b} \leftarrow std(B)$;
14 # Standard deviations of A and B
15 $RUL \leftarrow RUL \cup \mathbb{E}[Weibull(\bar{a}, \bar{b})]$;
16 $Var \leftarrow Var \cup mean([\hat{a}, \hat{b}])$; # average between $\hat{a}, \hat{b}$
17 **end**
18 $f_1 \leftarrow RMSE(RUL, Y_{V'})$;
19 $f_2 \leftarrow mean(Var)$; # Average value of Var
20 **Return** f_1, f_2

5. Experimental Setup and Results

We are interested in investigating the existence and trade-offs between the RUL prediction error and the prediction uncertainty when using bi-objective HPO, and to examine the advantages that can be gained compared to using a single-objective variant. Furthermore, we show how the proposed method can be more user-centric compared to the current techniques. Datasets and experimental results are described in this section.

5.1. Data

In this study, we use the widely used C-MAPSS benchmark dataset (A. Saxena & K. Goebel, 2008). The dataset was released in 2008 (Saxena et al., 2008) and it has been used in the field of PHM ever since, to develop techniques and methods for estimating the RUL (Ramasso & Saxena, 2014; Krishna &

Baghaei, 2019). It is a simulated turbofan engine degradation dataset from NASA's Prognostics Centre of Excellence[4]. The dataset consists of four subsets: FD001, FD002, FD003, and FD004, each of each exhibits a different number of operating conditions and fault modes. In this work, we used datasets FD001 and FD003, which exhibited the same number of operating conditions but different number of fault modes. Each of these datasets is arranged in an $n \times 26$ matrix where n corresponds to the number of data points (samples) in each unit and 26 is the number of columns/features. Each row is a snapshot of data taken during a single operating time cycle. Regarding the 26 features, the 1st represents the engine number, the 2nd represents the operational cycle number. Features $3-5$ represent the operational settings, and columns $6-26$ represent the 21 sensor values. Engine performance can be significantly affected by the three operating settings. More information about these 21 sensors can be found in (Ordóñez, Sánchez Lasheras, Roca-Pardiñas, & Juez, 2019). What is more, each subset exhibits a different number of faults (see Table 1).

Each of these subsets are further split into training set and test set (see Table 1 for details). For each engine trajectory within the training sets, the last data entry corresponds to the end-of-life (EoL) of the engine, i.e., the moment the engine is declared unhealthy or in failure status. The test sets contain data up to some time before the failure and the aim here is to predict the RUL for each of the test engines.

These multivariate time-series are from a different engine i.e., the data can be considered to be from a fleet of engines, of the same type though, and each trajectory is assumed to be the life-cycle of an engine. Every engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user. This wear and variation is considered normal, i.e., it is not considered a fault condition.

To compare the model performance on the test data, we need some objective performance measures. In this study, we us the *Root Mean Square Error (RMSE)* (Zheng et al., 2017; Listou Ellefsen et al., 2019; X. Li, Ding, & Sun, 2018), defined as: $RMSE = \sqrt{1/n \sum_{i=1}^{n} d_i^2}$, where $d_i = \hat{RUL}_i - RUL_i$, $\hat{RUL}_i$ is the estimated RUL and RUL_i is the ground truth RUL for instance (engine) i, respectively.

5.2. Experimental Setup

The experiments[5] were executed on 10 *NVIDIA Tesla T4* GPUs, of 16GB, *GDDR6* memory. Source code has been developed in `Python` V3.8.8[6]. Experimental time was around 3-5 days (wall clock time), per dataset.

[4] `https://ti.arc.nasa.gov/tech/dash/groups/pcoe/`

[5] The source code of the experiments can be found at `https://github.com/MariosKef/RULe`.

[6] We used `tensorflow`(2.5.0), `scikit-learn`(0.24.1), `pandas`(1.2.3), `numpy`(1.19.5).

Table 1. FD001 and FD003 C-MAPSS dataset details

Data-Set	FD001	FD003
Train trajectories	100	100
Test trajectories	100	100
Operating conditions	1	1
Fault conditions	1	2
Max train trajectory (cycles)	362	525
Min train trajectory (cycles)	128	145
Max test trajectory (cycles)	303	475
Min test trajectory (cycles)	31	38
Training samples	20631	24720

We began by randomly selecting 80% of units from the training set and using the remaining 20% as the validation set to select the hyperparameters. We then randomly truncate the trajectories of the validation set at five different locations such that five different cases are obtained from each trajectory following (Malhotra et al., 2016). The truncation is needed to replicate the dedicated test data, i.e., trajectories up to some time before the failure. Note here, however, that we did not use any information from the dedicated test set. Minimum truncation is 5% of the total life, and maximum truncation is 96% of the total life. We continued with the pre-processing of the training and validation sets. In more detail, we normalized the data transforming the 3 operational settings and 21 sensor values to the range $[-1, 1]$ (min-max normalization) and discarded any of them that have zero variance. Constant values do not provide any useful degradation information for determining the RUL.

For the next steps of the pre-processing and data transformation (sliding window and RUL target construction), as well as for the DNN training, we performed HPO to select their optimal hyperparameter values that optimize *simultaneously* the pointwise RMSE and the uncertainty, in order to address our 1st and 2nd contributions (see Section 4.5). The tuned hyperparameters and their respective ranges can be seen in Table 2. Note that the search space contains not only integer variables but also categorical ones. We executed the hyperparameter optimization (see Section 4.5) with a budget of 300 function evaluations (of which 100 are initial configurations sampled with the latin hypercube sampling (LHS) method). Moreover, the MIP-EGO configurator is set to evaluate 10 configurations per step in parallel for FD001 and 9 configurations for dataset FD003[7].

Following the hyperparameter optimization phase, we are presented with a two-dimensional set of points showing the RMSE and UQ on the *validation set*. Each point corresponds to a specific hyperparameter configuration. By considering only the non-dominated solutions, we end up with (an approximation to) the Pareto front. The Pareto front is set of points, which cannot be improved with respect to one objective without making another objective worse (Emmerich & Deutz,

[7] This was a result of GPU availability. In any case, this did not affect the validity of the computations.

2018) (see blue points in Figure 2). The non-dominated set of solutions delivers hyperparameter configurations which allow us to view the trade-offs between the RMSE and the UQ. We can subsequently pre-process and train on the *entirety* of the training data (training and validation) using the configurations corresponding to the points on the Pareto front and finally test our method on the dedicated test set. During this stage, we use Algorithm 1 by inputting as X the entire training set, V the dedicated test set, and h_p the configuration corresponding to the selected point from the Pareto front.

Additionally, we used the Adam optimizer (Kingma & Ba, 2017) with a clip value of 0.5, $R = 30$ for the number of MC Dropout passes, and trained for 100 epochs with early-stopping (patience = 5). Finally, since we want our DNN to learn the relationship between the input sequences and the Weibull parameters, we used as a loss function the *negative log-likelihood of the* 2*-parameter Weibull distribution* (Yang, Ren, & Hu, 2019; Martinsson, 2016) to train the network.

5.2.1. Baseline

We also performed a baseline experiment to evaluate the bi-objective hyperparameter approach. Our baseline differs from the work we reviewed in Section 2, as none of the related work took into account the joint optimization of the RMSE and the uncertainty. Our baseline transforms the bi-objective optimization problem into a single-objective by minimizing the harmonic mean (HM) of the RMSE and uncertainty, as:

$$HM = \frac{2}{RMSE^{-1} + Uncertainty^{-1}} \tag{6}$$

For this task we used the single-objective MIP-EGO, which uses the so-called Moment-Generating Function (MGF) based infill-criterion (H. Wang, van Stein, Emmerich, & Back, 2017) to select new candidate solutions. Moreover, the MIP-EGO configurator is set to evaluate 10 configurations per step in parallel for FD001 and FD003, for a maximum of 300 function evaluations. We used this baseline in order to investigate the benefits of using the bi-objective HPO compared to the single-objective approach. The reason of taking the HM compared to e.g., the arithmetic mean, is because it is less susceptible to fluctuation of the observations, thus making it a more ideal baseline for this first study.

5.3. Hypervolume Indicator

To compare the bi-objective HPO approach to the single-objective approach based on the HM we decided to use the hypervolume indicator (HVI). The HVI or S-metric (Zitzler, Deb, & Thiele, 2000) is the hypervolume in the objective space $\mathbb{R}^m$ that is dominated by the Pareto points bounded by a reference point $y_{ref} \in \mathbb{R}^m$. The reason for choosing the HVI as a measure of comparison is that it is intuitive, as dominating a large part of the objective space is desirable. Furthermore, the HVI is widely used in evaluating the performance of various multi-objective optimization algorithms.

5.4. Results and Discussion

Having generated the Pareto front of the hyperparameter configurations (see Section 5.2) we selected each configuration, trained on the entirety of the dataset and made inferences about both the training and (dedicated) test data.

Figures 2 and 3 show in blue circles the Pareto front of the hyperparameter configurations performance on the *validation sets* of datasets FD001 and FD003, respectively. The red triangles depict the results on the dedicated test set (dominated solutions might exist). The number next to each point represents the hyperparameter configuration giving rise to that specific solution and are shown here to manifest how the solutions' topology changes when validated on the dedicated test set.

In order to see if the neural network can learn from the data, in Figures 4 and 5, we show the evolution, over time, of the Weibull PDFs, of units 2 and 9 from the FD001 and FD003 training data, respectively. We do this by plotting the Weibull PDFs per time-index of the units' data. For this task, we used the models which returned the lowest RMSE on the *dedicated test sets* of FD001 and FD003 (points with green shade in Figures 2 and 3). In the Figures, we can see that as the time-index of the data increases (darker-red shades in the legend), the PDFs variance decreases. Even though the distributions' variance does not initially seem to be monotonically decreasing, as we approach the end-of-life of the assets (darker-red shades), we can see that the variance decreases, giving more mass to the expected time-to-failure, and that the expected time-to-failure approaches 0. This is a desirable property as it indicates that the model can learn the correct *failure dynamics* because the more time-steps have passed, the more data has been collected, and consequently, there is more degradation information, especially near the end-of-life of the asset.

In Figures 6 and 7 we show the evolution of the HVI per a maximum of 300 function evaluations between the bi-objective and single-objective HPO. To be able to compare the HVI of the single-objective approach to the bi-objective approach, we calculated the HVI of the Pareto efficient solutions of the RMSE and uncertainty as pre-images of the HM. Furthermore, we normalized both objectives to $[0, 1]$ and used as $y_{ref} = (1.1, 1.1)$.

We can see from the two figures that the HVI of the single-objective approach and the bi-objective approach plateau to the same final HVI, albeit the bi-objective approach reaches the plateau in fewer iterations, on FD001, whereas on FD003, the single-objective approach reaches the plateau in slightly fewer iterations than the bi-objective method. The HVI might indicate that the harmonic mean manages to also identify a

Table 2. Hyperparameters in the model development for the C-MAPSS dataset

Type	Hyperparameter	Search Space
Pre-processing	Sliding window size	[20, 50]
	Reflection point (percentage of total life)	[25, 75]
	Initial RUL value	[110, 130]
	RUL degradation style	['linear', 'nonlinear']
DNN	Number of recurrent layers	[1, 3]
	Number of dense layers	[1, 3]
	Number of neurons per layer	[10, 100]
	Activations	['tanh', 'sigmoid']
	Recurrent dropout rate	[1e-5, 0.9]
	Dropout rate	[1e-5, 0.9]
	Output activations	['softplus', 'exp']
	Learning rate	[1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 2e-5]
	Batch size	[32,64,128]

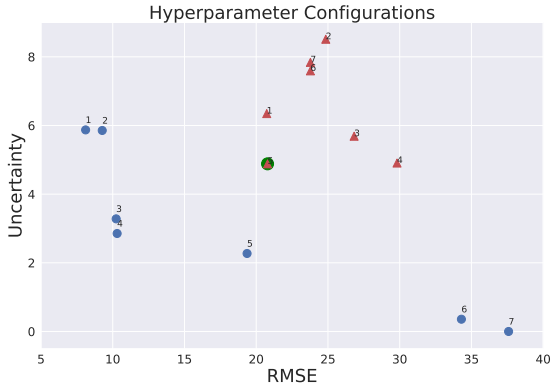

Figure 2. RMSE-UQ points corresponding to the hyperparameter configurations on FD001 using the bi-objective approach. Blue circles are the Pareto front as calculated on the validation set. The red triangles are the points calculated on the dedicated test set.

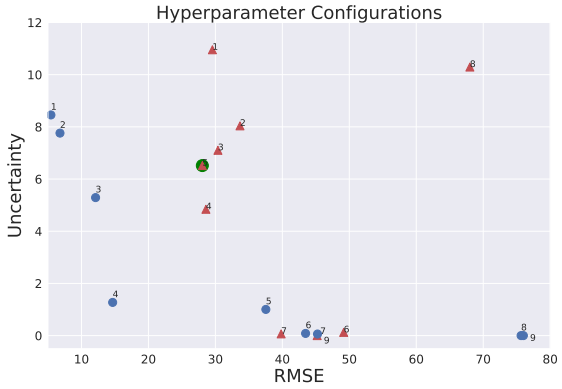

Figure 3. RMSE-UQ points corresponding to the hyperparameter configurations on FD003 using the bi-objective approach. Blue circles are the Pareto front as calculated on the validation set. The red triangles are the points calculated on the dedicated test set.

balance between the objectives and can be used as an alternative to the bi-objective HPO. The seemingly smaller number of function evaluations of the single-objective approach in the figures, compared to the bi-objective approach, is simply an artifact of infeasible configurations that were discarded by the single-objective MIP-EGO.

Examining Figures 2 and 8 we can see that the bi-objective approach returned more hyperparameter configurations lying on the Pareto front (7 blue points on Figure 2) compared to the single-objective approach (6 blue points on Figure 8). Even though the number is marginally larger, this might suggest that the bi-objective approach might be more suitable for identifying a more diverse set of hyperparameters. Moreover, it is interesting to see that the configurations returned from the two HPO methods (blue points in Figures 2 and 8) present similar values of uncertainty, even though more than 80% of the configurations of the single-objective HPO exhibit uncertainty lower than 2, with that number being around 29% for the bi-objective HPO. Regarding RMSE, however, we observe the inverse trend. In the bi-objective method, more than 70% of the returned configurations result in RMSE lower than 20, with this number being 50% in the single objective approach. In addition, we can see that the performance of the resulting hyperparameters (blue points) on the dedicated test set (red triangles) differs between the two figures. Firstly, in the bi-objective approach, the performances on the dedicated test set per hyperparameter configuration are clustered together when compared to the single-objective approach in Figure 8 where the points are spread out more, especially in the uncertainty axis. Secondly, in the bi-objective method, the RMSE and uncertainty values of the dedicated test set lie in the range of $[20.73, 29.82]$ and $[4.88, 8.51]$, respectively. In the single-objective method these ranges are $[25.97, 37.51]$ and $[0, 7.93]$, respectively, for the RMSE and uncertainty. It is interesting

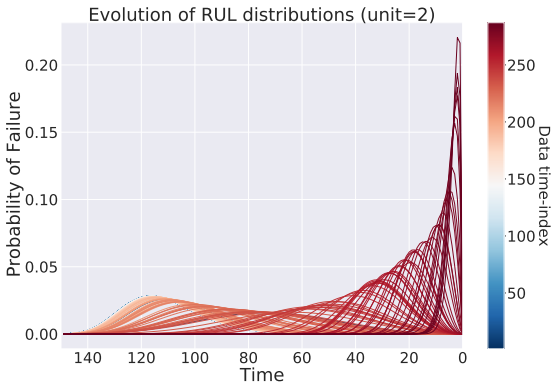

Figure 4. Evolution of Weibull distributions of unit 2 from FD001. Blue shades indicate the start of the unit's trajectory and red shades the end. Note that the x-axis is inverted for clarity.

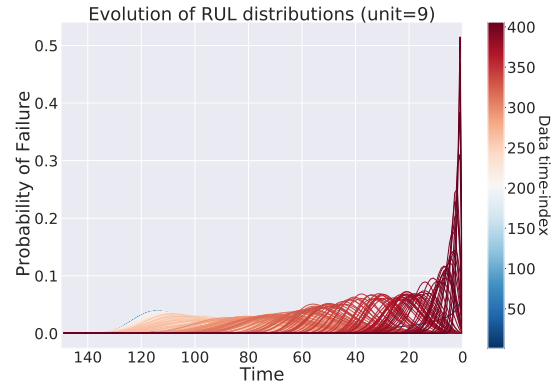

Figure 5. Evolution of Weibull distributions of unit 9 from FD003. Blue shades indicate the start of the unit's trajectory and red shades the end. Note that the x-axis is inverted for clarity.

to see that the bi-objective HPO returned better scores for the RMSE and more "concentrated scores" for the uncertainty compared to the single-objective approach.

Regarding FD003 when examining Figures 3 and 9 we can see that the bi-objective approach returned, again, a larger number of hyperparameter configurations lying on the Pareto front (9 blue points on Figure 3) compared to the single-objective approach (7 blue points on Figure 9). Even though the number is marginally larger, this suggests, like previously, that the bi-objective approach might be more suitable for identifying a more diverse set of hyperparameters. In the bi-objective method, around 44% of the returned configurations result in RMSE lower than 20, with this number being around 57% in the single-objective approach. Nevertheless, we observe that the hyperparameter configurations from the bi-objective approach returned overall configurations with lower levels of uncertainty compared to the single-objective method. Specifically, more than 66% of the configurations on the bi-objective HPO result in uncertainty that is less than 2, with this number being around 43% in the single-objective HPO. Regarding the resulting hyperparameters' performance (blue points) on the dedicated test set (red triangles), there are no apparent differences between the two methods' topologies. Lastly, in the bi-objective method, the RMSE and uncertainty values of the dedicated test set lie in the range of $[28.05, 68.01]$ and $[0, 10.96]$, respectively. In the single-objective method, these ranges are $[23.82, 50.76]$ and $[0.14, 18.53]$, respectively, for the RMSE and uncertainty. This shows that for this dataset, the bi-objective method returned lower uncertainty values, but the single-objective approach returned RMSE values that lie in a more favorable range, thus indicating no clear winner.

From the previous results, we conclude that the usage of bi-objective HPO can reveal interesting trade-offs between the RMSE and uncertainty. Additionally, the results show that even though the bi-objective approach can return more configurations on the Pareto front, the single-objective HPO is also a good alternative for this task. The differences in the experimental findings between the two datasets might be justified by the the fact that FD003 has 2 simulated fault conditions compared to FD001. In addition, we cannot rule out that the maximum allowable number of function evaluations or training epochs might have affected the findings, as more epochs might allow the network to learn more. More function evaluations of the HPO, on the other hand, will explore a larger part of the hyperparameter configuration space which might uncover better configurations.

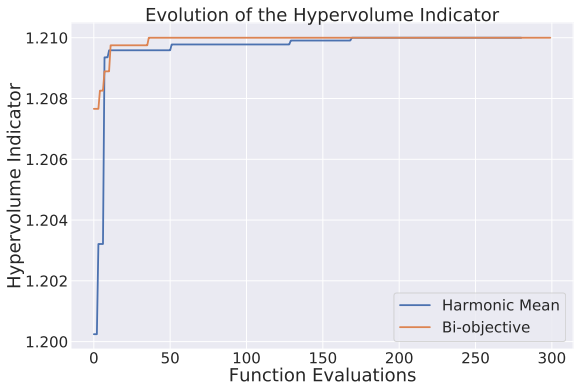

Figure 6. Evolution of the HVI of the bi-objective HPO and the single-objective HPO on FD001.

5.5. Application

Next, we will demonstrate how the proposed method can allow a more user-centric and interpretable approach to end-users (3rd contribution). For this application, we used the models which returned the lowest RMSE on the *dedicated test sets* of FD001 and FD003. These points are indicated with a green

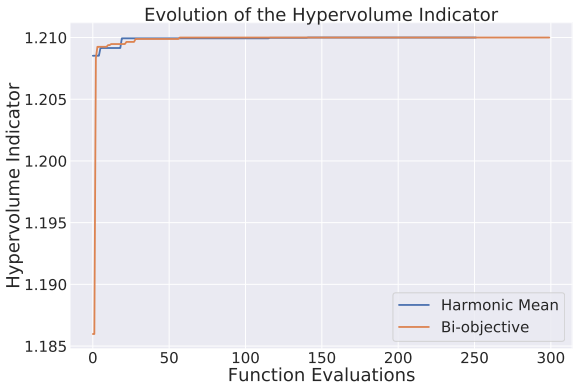

Figure 7. Evolution of the HVI of the bi-objective HPO and the single-objective HPO on FD003.

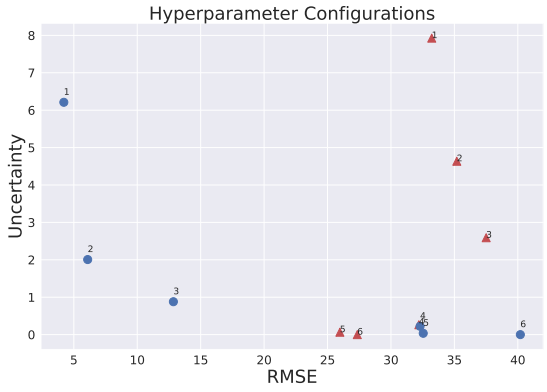

Figure 8. RMSE-UQ points corresponding to the hyperparameter configurations on FD001 using the harmonic mean approach. Blue circles are the Pareto front as calculated on the validation set. The red triangles are the points calculated on the dedicated test set.

marker on Figures 2 and 3. Specifically, since the trained network outputs the α and β parameters per input sample, the end-user can utilize this information to visualize, for example, the survival curves corresponding to each input sample, as well as other important information.

Survival curves are visualization methods from survival analysis that show the probability of an event *not* happening up to a point in time. In our case, this means that a failure has *not* occurred up to a point it time t (hence the asset will *survive* longer than t). A survival curve is defined as $1-$CDF, where CDF stands for the cumulative distribution function (in this case, the Weibull's CDF). For example in Figures 10 and 11 we plot the survival curves of test units $81, 4$ from the FD001 dataset and test units $28, 3$ from the FD003 dataset. For each test unit, we plot *all* the survival curves (shown within shaded areas for clarity) resulting from the multiple values of α and β that the network outputs through the MC Dropout, as well as the "median" curves that have as parameters the median values of the αs and βs, for a reference. This allows two things: the end-user can visually inspect the survival curves and, for instance, select a probability-of-survival threshold, based on one of them (e.g., the "median" curve), after which a unit should be maintained. Additionally, based on how wide the shaded areas are, the user can decide whether to employ the recommendation or proceed to further actions, such as further inspection by a field expert. For example, in Figure 11 the "median" survival curve of test unit 28 tells us that the probability of not having a failure up to time 100 from the current point in time (time 0) is about 80% and that this estimation is "more confident" compared to that of test unit 3, as the shaded area is less wide than the shaded area of test unit 3. Similarly, in Figure 10 the estimation of the survival curves of test unit 81 is "more confident" compared to that of test unit 4.

Figure 9. RMSE-UQ points corresponding to the hyperparameter configurations on FD003 using the harmonic mean approach. Blue circles are the Pareto front as calculated on the validation set. The red triangles are the points calculated on the dedicated test set.

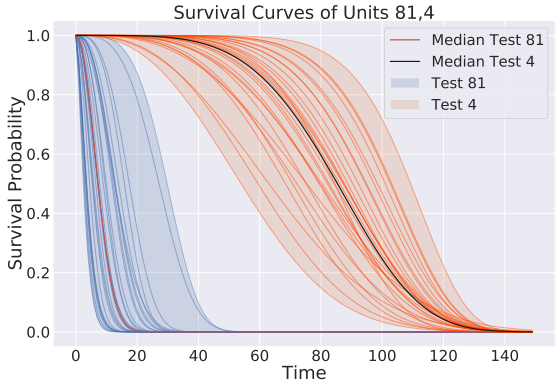

Figure 10. Survival curves of three units $81, 4$ from FD001. The shaded areas include *all* the survival curves from the multiple passes through MC Dropout.

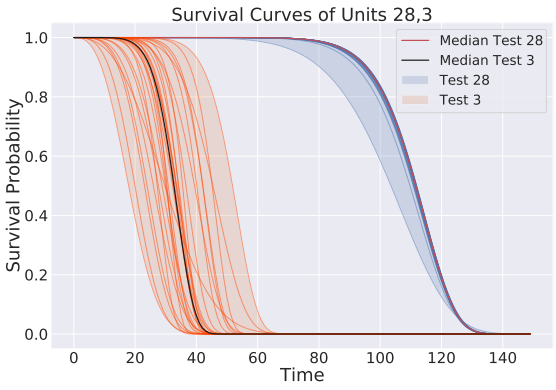

Figure 11. Survival curves of three units $28, 3$ from FD003. The shaded areas include *all* the survival curves from the multiple passes through MC Dropout.

6. Conclusions and Outlook

In this work, we dealt with the remaining useful life (RUL) estimation using Bayesian deep learning (BDL) by taking into consideration the uncertainty of the estimate together with the predicted point estimate. We investigated the first, to our knowledge, usage of *bi-objective* hyperparameter optimization (HPO) that minimizes *simultaneously* the pointwise RMSE and the uncertainty. In this direction, we optimized together with the hyperparameters of the neural network (NN) the hyperparameters that govern the pre-processing steps, delivering thus, an *end-to-end*, data-driven, pipeline for the (offline) RUL estimation. We validated our approach on two subsets of the famous C-MAPSS dataset (A. Saxena & K. Goebel, 2008). We, further, demonstrated how survival curves can provide the end-user with information regarding the RUL and its confidence.

The experimental results indicate that, the bi-objective HPO might be more suitable for identifying a more diverse set of hyperparameter configurations compared to the single-objective HPO that aggregates the two objectives through the harmonic mean (HM). However, both methods reach the same hypervolume indicator value of the Pareto front in, more or less, the same number of function evaluations and the findings did not indicate whether a method is more suitable for lower uncertainty or lower RMSE scores. Regarding the performance of the Pareto front configurations, when validated on the dedicated test sets, there was no clear winner between the two methods, although in the first examined case the RMSE values are better and the overall performance scores are clustered together. Overall, the results show that, for the examined cases, the bi-objective method is able to suggest more hyperparameter configurations and that the single-objective alternative is able to compete in terms of scores. This suggests that for a certain class of problems single-objective HPO methods are sufficient, allowing practitioners an ample selection of efficient single-objective HPO methods.

Concerning the limitations of our work, due to the high computational costs of running the experiments multiple times no statistical significance tests are performed. Despite that fact, our methodology is experimentally sound and suggests an alternative approach for HPO in PHM. Furthermore, as indicated, we are aware that there is a current debate as to the validity of Monte Carlo Dropout being Bayesian (Osband et al., 2018). This could, in turn, make the corresponding predictive models problematic in support of reliable uncertainty quantification. As this work was mainly devoted to the usage of bi-objective hyperparameter optimization and user-centric approach, we have decided to address this *highly relevant but challenging issue* in future work. Future work should, in general, emphasize research on computationally efficient and accurate uncertainty quantification of DL models, as this will further open the road of AI applied in real-world applications.

Finally, we would be very interested in extending the bi-objective HPO to a many-objective context to add more objectives, such as run-time, to find a compromise between accuracy, uncertainty, and training time. The authors hope that multi-objective hyperparameter optimization methods become a new alternative, as it is not the case that a single objective method can always capture the conflicting interests that exist in real-world problems.

Acknowledgment

This work is part of the research programme Smart Industry SI2016 with project name CIMPLO and project number 15465, which is partly financed by the Netherlands Organisation for Scientific Research (NWO).

References

A. Saxena, & K. Goebel. (2008). *Turbofan engine degradation simulation data set.* NASA Ames Research Center, Moffett Field.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., … Nahavandi, S. (2021, December). A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, *76*, 243–297. doi: 10.1016/j.inffus.2021.05.008

Benker, M., Furtner, L., Semm, T., & Zaeh, M. F. (2021, October). Utilizing uncertainty information in remaining useful life estimation via Bayesian neural networks and Hamiltonian Monte Carlo. *Journal of Manufacturing Systems*, *61*, 799–807. doi: 10.1016/j.jmsy.2020.11.005

Biggio, L., Wieland, A., Chao, M. A., Kastanis, I., & Fink, O. (2021, April). Uncertainty-aware Remaining Useful Life predictor. *arXiv:2104.03613 [cs, stat]*.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017, April). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. doi: 10.1080/01621459.2017.1285773

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015, May). Weight Uncertainty in Neural Networks. *arXiv:1505.05424 [cs, stat]*.

Caceres, J., Gonzalez, D., Zhou, T., & Droguett, E. L. (2021, October). A probabilistic Bayesian recurrent neural network for remaining useful life prognostics considering epistemic and aleatory uncertainties. *Structural Control and Health Monitoring*, *28*(10). doi: 10.1002/stc.2811

den Hertog, D., Kleijnen, J. P. C., & Siem, A. Y. D. (2006, April). The correct Kriging variance estimated by bootstrapping. *Journal of the Operational Research Society*, *57*(4), 400–409. (Publisher: Taylor & Francis _eprint: https://doi.org/10.1057/palgrave.jors.2601997) doi: 10.1057/palgrave.jors.2601997

Emmerich, M. T. M., & Deutz, A. H. (2018, September). A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing*, *17*(3), 585–609. doi: 10.1007/s11047-018-9685-y

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges* (pp. 3–33). Cham: Springer International Publishing. doi: 10.1007/978-3-030-05318-5_1

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd international conference on international conference on machine learning - volume 48* (p. 1050–1059). JMLR.org.

Goodfellow, I., Yoshua Bengio, & Aaron Courville. (2016). *Deep Learning*. MIT Press.

Govaers, F. (2019). *Introduction and Implementations of the Kalman Filter*. BoD – Books on Demand.

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*.

Hsu, C.-S., & Jiang, J.-R. (2018, April). Remaining useful life estimation using long short-term memory deep learning. In *2018 IEEE International Conference on Applied System Invention (ICASI)* (pp. 58–61). Chiba: IEEE. doi: 10.1109/ICASI.2018.8394326

Hüllermeier, E., & Waegeman, W. (2021, March). Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, *110*(3), 457–506. doi: 10.1007/s10994-021-05946-3

Kalman, R. E. (1960, March). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. doi: 10.1115/1.3662552

Kefalas, M., Baratchi, M., Apostolidis, A., van den Herik, D., & Bäck, T. (2021, June). Automated Machine Learning for Remaining Useful Life Estimation of Aircraft Engines. In *2021 IEEE International Conference on Prognostics and Health Management (ICPHM)* (pp. 1–9). Detroit (Romulus), MI, USA: IEEE. doi: 10.1109/ICPHM51084.2021.9486549

Kim, M., & Liu, K. (2021, March). A Bayesian deep learning framework for interval estimation of remaining useful life in complex systems by incorporating general degradation characteristics. *IISE Transactions*, *53*(3), 326–340. doi: 10.1080/24725854.2020.1766729

Kingma, D. P., & Ba, J. (2017, January). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kiureghian, A. D., & Ditlevsen, O. (2009, March). Aleatory or epistemic? Does it matter? *Structural Safety*, *31*(2), 105–112. doi: 10.1016/j.strusafe.2008.06.020

Kraus, M., & Feuerriegel, S. (2019, October). Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. *Decision Support Systems*, *125*, 113100. doi: 10.1016/j.dss.2019.113100

Krishna, M., & Baghaei, K. T. (2019). Recent Approaches in Prognostics: State of the Art. In *2019 International Conference on Artificial Intelligence (ICAI)* (pp. 358–365).

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018, may). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. doi: 10.1016/j.ymssp.2017.11.016

Li, G., Yang, L., Lee, C.-G., Wang, X., & Rong, M. (2021, September). A Bayesian Deep Learning RUL Framework Integrating Epistemic and Aleatoric Uncertainties. *IEEE Transactions on Industrial Electronics*, *68*(9), 8829–8841. doi: 10.1109/TIE.2020.3009593

Li, X., Ding, Q., & Sun, J.-Q. (2018, April). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, *172*, 1–11. doi: 10.1016/j.ress.2017.11.021

Listou Ellefsen, A., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019, March). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Reliability Engineering & System Safety*, *183*, 240–251. doi: 10.1016/j.ress.2018.11.027

Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016, August). Multi-Sensor Prognostics using an Unsupervised Health Index based on LSTM Encoder-Decoder. *arXiv:1608.06154 [cs]*.

Martinsson, E. (2016). *Wtte-rnn: Weibull time to event recurrent neural network* (Unpublished master's thesis). University of Gothenburg, Sweden.

Nguyen, V. D., Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., & Bäck, T. (2019). A Review: Prognostics and Health Management in Automotive and Aerospace. *International Journal of Prognostics*

and Health Management, *10*(2), 35. doi: 10.36001/ijphm.2019.v10i2.2730

Ordóñez, C., Sánchez Lasheras, F., Roca-Pardiñas, J., & Juez, F. J. d. C. (2019, January). A hybrid ARIMA–SVM model for the study of the remaining useful life of aircraft engines. *Journal of Computational and Applied Mathematics*, *346*, 184–191. doi: 10.1016/j.cam.2018.07.008

Osband, I., Aslanides, J., & Cassirer, A. (2018, November). Randomized Prior Functions for Deep Reinforcement Learning. *arXiv:1806.03335 [cs, stat]*.

Peng, W., Ye, Z.-S., & Chen, N. (2020, March). Bayesian Deep-Learning-Based Health Prognostics Toward Prognostics Uncertainty. *IEEE Transactions on Industrial Electronics*, *67*(3), 2283–2293. doi: 10.1109/TIE.2019.2907440

Ramasso, E., & Saxena, A. (2014). Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets. *International Journal of Prognostics and Health Management*, *5*(2), 15. doi: https://doi.org/10.36001/ijphm.2014.v5i2.2236

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press.

Sateesh Babu, G., Zhao, P., & Li, X.-L. (2016a). Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. In S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. S. Wang, & H. Xiong (Eds.), *Database Systems for Advanced Applications* (Vol. 9642, pp. 214–228). Cham: Springer International Publishing. (Series Title: Lecture Notes in Computer Science) doi: 10.1007/978-3-319-32025-0_14

Sateesh Babu, G., Zhao, P., & Li, X.-L. (2016b). Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. In S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. S. Wang, & H. Xiong (Eds.), *Database Systems for Advanced Applications* (Vol. 9642, pp. 214–228). Cham: Springer International Publishing. (Series Title: Lecture Notes in Computer Science) doi: 10.1007/978-3-319-32025-0_14

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008, October). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management* (pp. 1–9). Denver, CO, USA: IEEE. doi: 10.1109/PHM.2008.4711414

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation – A review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1 – 14. doi: https://doi.org/10.1016/j.ejor.2010.11.018

Stein, B. v., Wang, H., & Back, T. (2019, July). Automatic Configuration of Deep Neural Networks with Parallel Efficient Global Optimization. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). Budapest, Hungary: IEEE. doi: 10.1109/IJCNN.2019.8851720

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014, February). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). Intelligent Fault Diagnosis and Prognosis for Engineering Systems.. doi: 10.1002/9780470117842

Wang, B., Lei, Y., Yan, T., Li, N., & Guo, L. (2020, February). Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery. *Neurocomputing*, *379*, 117-129. doi: https://doi.org/10.1016/j.neucom.2019.10.064

Wang, H., van Stein, B., Emmerich, M., & Back, T. (2017, October). A new acquisition function for Bayesian optimization based on the moment-generating function. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 507–512). Banff, AB: IEEE. doi: 10.1109/SMC.2017.8122656

Yang, F., Ren, H., & Hu, Z. (2019, May). Maximum Likelihood Estimation for Three-Parameter Weibull Distribution Using Evolutionary Strategy. *Mathematical Problems in Engineering*, *2019*, 1–8. doi: 10.1155/2019/6281781

Zhang, C., Lim, P., Qin, A. K., & Tan, K. C. (2017, October). Multiobjective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2306–2318. doi: 10.1109/TNNLS.2016.2582798

Zhao, Z., Wu, J., Wong, D., Sun, C., & Yan, R. (2020). Probabilistic Remaining Useful Life Prediction Based on Deep Convolutional Neural Network. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3717738

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017, June). Long Short-Term Memory Network for Remaining Useful Life estimation. In *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)* (pp. 88–95). Dallas, TX, USA: IEEE. doi: 10.1109/ICPHM.2017.7998311

Zhou, T., Droguett, E. L., Mosleh, A., & Chan, F. T. S. (2021, October). An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications. *arXiv:2111.00874 [cs]*.

Zitzler, E., Deb, K., & Thiele, L. (2000, June). Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evolutionary Computation*, *8*(2), 173–195. doi: 10.1162/106365600568202

BIOGRAPHIES

Marios Kefalas currently pursues his Ph.D. in Predictive Maintenance and Optimization at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands. He received his BSc degree in Pure and Applied Mathematics at the Department of Mathematics, University of Athens, Greece, in 2015 and his MSc degree in Bioinformatics at LIACS, Leiden University, The Netherlands, in 2017. His research interests lie in Prognostics and Health Management, time-series application in industry and health, and Data Science.

Bas van Stein received his Ph.D. degree in Computer Science in 2018, from the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands. From 2018 until 2021 he was a Postdoctoral Researcher at LIACS, Leiden University and he is currently an Assistant Professor at LIACS. His research interests lie in surrogate assisted optimisation, surrogate assisted neural architecture search and explainable AI techniques for industrial applications.

Mitra Baratchi received her Ph.D. degree in Computer Science in 2015, from the University of Twente, The Netherlands. In 2017, she joined the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands, as an Assistant Professor. Her research interests lie in machine learning for spatio-temporal and time-series data targeting various environmental and industrial applications.

Asteris Apostolidis received his Ph.D. degree in Computational Aerothermodynamics in 2015 from Cranfield University, UK. He worked for aircraft manufacturers, airlines and academic institutes and he is currently appointed as an Associate Professor at Amsterdam University of Applied Sciences. His interests include physics-based and data-driven methods for aircraft systems simulation, aircraft Maintenance Repair and Overhaul (MRO) and novel propulsion architectures.

Thomas Bäck (Fellow, IEEE) received the Diploma degree in Computer Science in 1990 and the Ph.D. degree in Computer Science in 1994, both from the University of Dortmund, Germany. Since 2002, he is Full Professor of Computer Science with the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands. His research interests include evolutionary computation, machine learning, and their real-world applications, especially in sustainable smart industry and health.